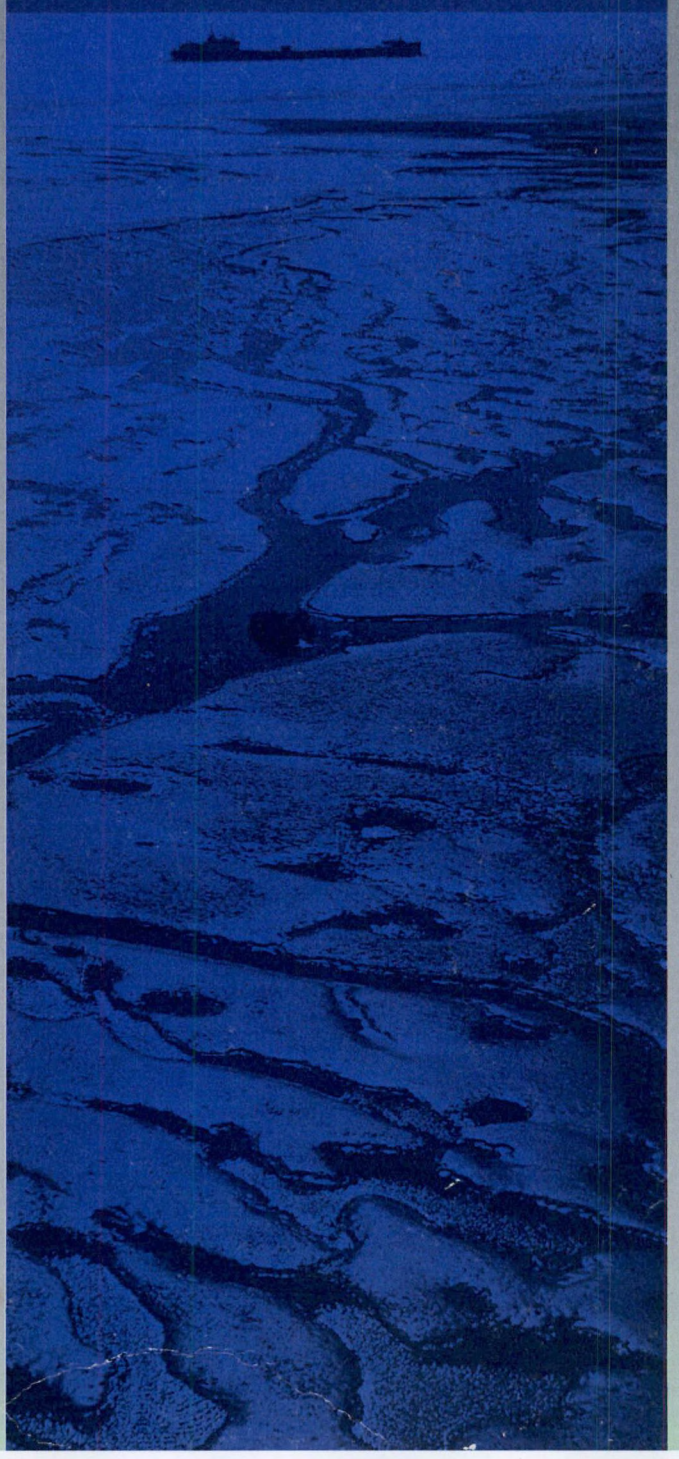
erkeer en waterstaat

35

rijkswaterstaat

dienst getijdewateren



M

J.L. Bruijs

INLEIDING TOT HET GEBRUIK VAN BESCHRIJVENDE MULTIVARIATE TECHNIEKEN VOOR DE VERWERKING VAN ECOLOGISCHE GEGEVENS.

Patrick Meire en Martin Hermy°

Rijksuniversiteit Gent
Laboratorium voor Ecologie der
Dieren, Zoögeografie en Natuurbehoud
Ledeganckstraat 35
B9000 Gent
België

°Ministerie van de Vlaamse
Gemeenschap
Instituut voor Natuurbehoud
Kiewitdreef 3
B3500 Hasselt
België

"Numerical methods are tools and unless they are used in the investigation of real ecological problems we are wasting our time in developing them" (Greig-Smith, 1980).

Eerste versie !!!

Het voorliggende rapport is een eerste versie en zal nog grondig herwerkt worden op basis van de te verwachten commentaar. In ieder geval moet de relatie tussen de diverse technieken nog beter uitgewerkt worden, moet dieper ingegaan worden op Gaussische responscurves en ordinatie en moet bij Canoco nog wat extra informatie komen. Ook Discrim komt nog aan bod. In een appendix zullen ook een tweetal voorbeelden grondig uitgewerkt worden met de diverse methoden.

18.05.89

INHOUDSTAFEL

1) Inleiding 4

2) Overzicht van de verschillende fazen in het onderzoek 5

3) Het invoeren van de gegevens 7
kenmerken van een dataset 7
volle versus gecondenseerde matrix 8
freefield versus fixed format 9
fortran formating 10

4) Schema van de gegevensverwerking 12

5) Aanpassen van de datamatrix 13
Transformatie 13
Standardizeren en relativeren 14
Z-scores 16
Logaritmeren 17
Vierkantswortel transformatie 18
Box-Cox transformatie 18
Arcussinus transformatie 18
Binaire transformatie 19
groeperen in klassen 19
vergelijking van enkele transformaties 19
Datareductie 21

6) Ordinatie 23
Principal Component Analysis 26
Geometrische kenmerken van de dataset 27
Geometrische structuur van een correlatiematrix 28
Hoofdcomponenten-transformatie 30
Opname ordinatie 33
R- en Q- ordinaties 33
Berekening via matrix algebra 33
Diverse PCA varianten 33
Multidimensional Scaling 36
Weighted Average ordinatie 39
Reciprocal Averaging (Correspondance analysis) .. 39
Detrended Correspondance Analysis 41
problemen met RA 41
mogelijke oplossingen 44
testen en problemen met DCA 47
enkele belangrijke programma-specificaties . 48
uitbijters 48
transformaties 48
downweighting 48
herschaling van de assen een segmenten . 49
Het ordinatie resultaat 49
Canonical Correspondance Analysis 50
Verdere uitwerking van CANOCO 54

7) Klassificatie of Clusteranalyse 57
Similariteitsindices 57
Sorensen 58
Renkonen 59

Bray-Curtis 59
Canberra metric 59
Czekanovski 60
Similarity ratio 60
Euclidische aftsand 60
enkele kenmerken van sommige indices 61
Sorteringsmethoden 62
Nearest Neighbour Sorting 63
Furthest Neighbour Sorting 63
Group Average Sorting 64
Ward's analyse 65
Flexible Sorting 65
Een hybride klassificatie techniek: TWINSPAN 67
Pseudosoorten en cutlevels 67
Maken van dichotomieen 68
Monster klassificatie 68
Ordenen van de dichotomieën 71
Soortsklassificatie 71
Two-way tabel 72
Enkele bedenkingen 73
DISCRIM 74

8) Diversiteitsmaten .....

9) Een volledig uitgewerkt voorbeeld .....

10) Conclusie .....

11) Dankwoord 76

12) Literatuur 76

Appendix 1: Overzicht van de besproken programma's ....

Appendix 2: Handleiding van het programma CONDENS ....

Appendix 3:

1) INLEIDING

In de ecologie kunnen wij ruwweg twee typen onderzoek onderscheiden: experimenteel onderzoek waar een specifieke hypothese of in het veld of in het labo getoetst wordt en beschrijvend vergelijkend onderzoek waar door het bemonsteren van een gebied gepoogd wordt inzicht te verkrijgen in het voorkomen van organismen en de factoren die hiervoor van belang zijn. De inherente complexiteit van deze laatste tegenover de elegantie van een experimentele benadering op een Popperiaanse wijze heeft soms tot gevolg gehad dat het beschrijvend onderzoek als minderwaardig werd beschouwd en bijgevolg minder aandacht kreeg. Toch schrijft Pianka (1987): "In my opinion, of the various subdisciplines in ecology, the study of communities is the most abstract and most tantalizing, most important and most urgent, but unfortunately also the most difficult. We need to improve our understanding of how ecological systems behave. The traditional pair-wise approach to population interactions has proven totally inadequate. We must attempt to understand interactions within complex networks.". Gezien de complexiteit van het geheel zijn we in een beschrijvende benadering grotendeels aangewezen op statistische technieken.

In recente jaren zijn diverse "nieuwe" verwerkingsmethoden ontwikkeld die ons precies kunnen helpen om op basis van grote hoeveelheden monsters inzicht te krijgen in de structuur van het ecosysteem en in de belangrijkste onderliggende relaties, de zogenoemde multivariate methoden.

Een entiteit (een monster of een soort) wordt gekarakteriseerd door meerdere kenmerken (bijvoorbeeld door diverse soorten en abiotische factoren of door diverse morfologische kenmerken) zodat we dus met MULTIVARIABELE ENTITEITEN te doen hebben. Bovendien zijn vele variabelen onderling gecorreleerd, zodat er in feite een deel redundante informatie in elke dataset vervat zit (Krzanowski,1972 in Gauch,1982) :"The need for multivariate analysis arises whenever more than one characteristic is measured on a number of individuals, and relationships among the characteristics make it necessary for them to be studied simultaneously.".

De multivariate methoden kunnen we grofweg indelen in twee types: technieken die ons toelaten om bepaalde patronen, in de dataset aanwezig, op te sporen en te beschrijven en anderzijds technieken die specifiek gebruikt kunnen worden om bepaalde hypothesen te toetsen. De beschrijvende technieken kunnen we verder opdelen in 2 grote groepen: classifcatie en ordinatie.

Het zijn precies die methoden die waarvan we in dit rapport een overzicht en een bechrijving willen geven. Daarnaast willen we ook een soort standaard procedure voorschrijven die we kunnen volgen bij de analyse van ecologische data. Belangrijk hierbij is dat we proberen die informatie te geven die nodig is om bij het uitvoeren van de programma's een gegronde keuze te maken tussen de diverse geboden alternatieven en de gehele output op de listings te kunnen interpreteren. Meer gedetailleerde informatie over deze beschrijvende multivariate methoden kunnen we vinden in Clifford & Stephenson (1975), Gauch (1982), Greig-Smith (1983), Jongman et al. (1987), Legendre & Legendre (1986) en Pielou (1984). Hypothese testende multivariate methoden worden beschreven in (referenties). Een overzicht van diverse methoden

van een post-hoc interpretatie van gegevens is gegeven in Heip et al. (1988). Belangrijke handboeken voor parametrische statistiek zijn Snedecor & Cochran (1980), Sokal & Rholf (1981), voor de niet-parametrische statistiek Conover (1980) en Siegel (1956).

2) OVERZICHT VAN DE VERSCHILLENDE FAZEN IN HET ONDERZOEK

Een studie van bepaalde levensgemeenschappen, biotopen etc. bestaat in principe uit drie verschillende onderdelen (naar Hermy, 1984): een verzamelings fase waarin de gegevens samengebracht worden, een verwerkings fase waarin de verzamelde data geanalyseerd worden, en een interpretatie fase waarin de resultaten binnen bestaande theoretische denkkaders of ecologische hypothesen worden ingepast. Deze verschillende fazen zijn weergeven in Fig. 2.1. Hier zijn eveneens de in dit rapport besproken programma's vermeld. In Appendix 1 wordt technische informatie over de diverse programma's gegeven.

Het begin van de data verzamelings fase en van elk onderzoek is de omschrijving en de afbakening van de doelstellingen. Een duidelijke omlijning van deze doelstellingen kan de analyse en de verwerking aanzienlijk vergemakkelijken. Het is bovendien van essentieel belang bij de keuze hoe de data in het veld moeten verzameld worden. Hoewel we in het kader van dit rapport daar niet nader kunnen op ingaan moeten we hier toch enkele algemeenheden vermelden. Afhankelijk van wat we willen onderzoeken is het nodig om een goed beeld te krijgen van het aanwezige soortenspectrum of om van een aantal soorten een goede dichtheidsschatting te maken. Beide zijn zeer sterk afhankelijk van enerzijds het bemonsterde oppervlak en anderzijds van het aantal genomen monsters. Het verdient dan ook aanbeveling om eerst een soort pilootbemonstering uit te voeren om een idee te krijgen van het soortenspectrum en de ruimtelijke spreiding van de organismen. Op basis daarvan kan dan de eigenlijke bemonstering gebeuren. Voor verder informatie verwijzen we naar Cochran (1967), Snedecor & Cochran (1981),.. We kunnen niet genoeg benadrukken dat al het verdere werk staat of valt met een goede bemonstering. In vele gevallen willen we ook een relatie nagaan met abiotische factoren. Ook aan het verzamelen van deze gegevens moet voldoende aandacht besteed worden. Eens de gegevens verzameld, worden ze in de gechikte vorm gebracht voor verdere verwerking. De computerinvoer gebeurt via een editor programma en de ingevoerde gegevens kunnen via het programma CONTROLDATA of via de editor gecheckt en verbeterd worden.

Na controle van de verzamelde gegevens moeten de data in functie van de doelstellingen verwerkt worden. In eerste instantie is het nuttig om een algehele dataverkenning uit te voeren aan de hand van simpele basisstatistiek. Dit omvat onder andere het nagaan van de minima en maxima per soort, maken van frequentiedistributies van de aantallen van enkele belangrijke soorten. Uit dit laatste kunnen we de verdeling nagaan, wat verder een basis voor de keuze van een transformatie kan zijn. Dit kan gebeuren via diverse statistiek programma's zoals SPSS, BMDP, SYSTAT, BIOM, STATGRAPH of via het Cornell Programma DATAEDIT. (In de Cornell University, Ithaca, New York werden binnen de afdeling plantenecologie, waar de zeer bekende ecoloog Whittaker werkte, een hele serie methoden en programma's ontwikkeld voor de analyse van gegevens zoals bv. TWINSPAN en

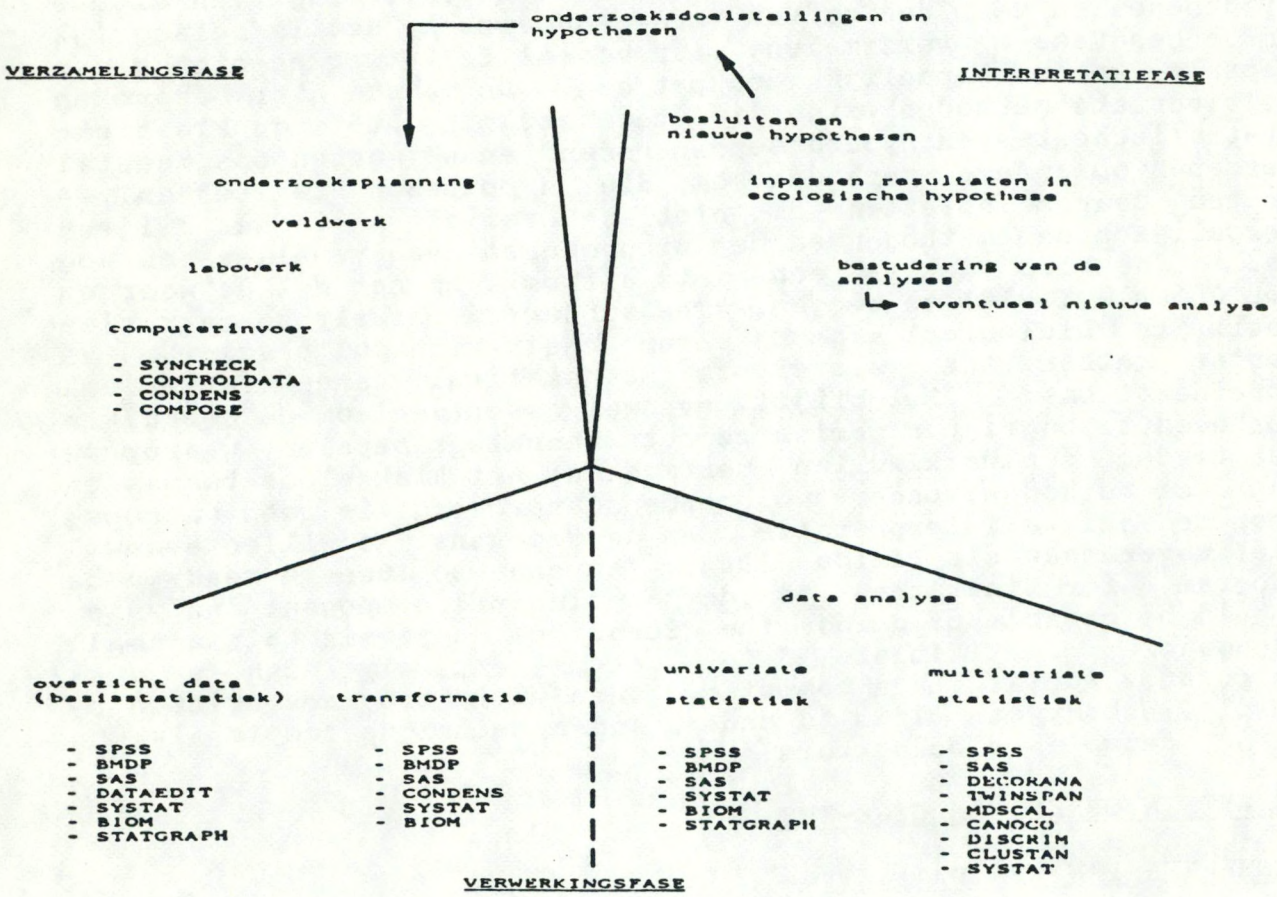


Fig. 2.1. Overzicht van de belanrijkste stappen in de loop van een onderzoeksproject, met inpassing van de in dit rapport besproken programma's.

DCA. In de verdere tekst zal nog gerefereerd worden naar Cornell of de CEP (Cornell Ecology Programs)). In tweede instantie, voor de eigenlijke data analyse, komt het behandelen en/of transformeren van de data. Dit kan ook weer via een hele reeks programma's o.a. de hoger genoemde evenals CONDENS, een CEP programma dat behoorlijk werd uitgebreid. Bij de data analyse kunnen we zowel univariate (parametrische of niet parametrische) statistiek (met SPSS, BIOM, BMDP, SAS, SYSTAT etc.) als diverse multivariate statistische methoden gebruiken (met SPSS, SYSTAT, TWINSPAN, DECORANA, CLUSTAN, CANOCO etc.).

Eens alle resultaten van de statistische analyses voorhanden zijn kunnen we eindelijk de resultaten interpreteren en toetsen aan de uitgangshypothese. Dit moet dan op zijn beurt leiden tot nieuwe analyses van de data of tot de formulering van nieuwe hypothesen en de planning van verder onderzoek. Het is belangrijk om de gegevens te verzamelen, maar vooral te verwerken binnen een bepaald wetenschappelijk denkpatroon. De meeste hier besproken multivariate methoden zijn in tegenstelling tot de klassieke statistische testen hypothese genererend en we moeten ons meestal beroepen op andere methoden om die hypothesen te testen. We moeten daarom opletten om niet te vervallen in het blijven vergelijken van methoden en het uitproberen van vanalles en nog wat in de hoop dat er eens iets uitkomt. Om het met de woorden van Goodall (1970) te zeggen: "The subject is likely to mark time until its biological aspects resume their rightful place i.e. as master rather than slave to mathematical and statistical considerations." Het blijft evenwel essentieel om de gebruikte methoden te begrijpen gezien ze het resultaat bepalen waarop we het verder denkwerk zullen verrichten. Het klakkeloos toepassen van deze methoden zonder voldoende achtergrond is geen waarborg voor de juiste interpretatie van de gegevens. We willen evenwel niet zover gaan als Pielou (1984): "Anyone who uses a ready-made program, for instance, to do a principal component analysis, should be capable of doing the identical analysis of a small manageable, artificial data matrix entirely with a desk calculator, or if on a computer, then with programs written by onself. Nobody can claim to understand a technique completely who is not capable of doing this."

3) HET INVOEREN VAN GEGEVENS

KENMERKEN VAN EEN DATASET.

Alle methoden vertrekken van gegevens die voorgesteld kunnen worden in tabellen met rijen (attributen, soorten, kenmerken) en kolommen (individuen, monsters..). Gemeenschappelijke kenmerken van data matrices zijn: ruis (noise), redundantie en uitbijters (outliers). Ruis verwijst naar het probleem dat replica's van monsters zelden identiek zijn. Hun verschillen kunnen het gevolg zijn van metingsverschillen, voor de waarnemer onzichtbare micro-verschillen in milieu, uniciteit van milieuomstandigheden, toevalsfactoren, etc.. Gegegevens bevatten bijgevolg deels interessante structuur en deels ruis. Onder redundantie verstaan we een grote mate van correlatie tussen de variabelen. Wanneer bv. meerdere soorten precies hetzelfde patroon van voorkomen vertonen dan kan dezelfde informatie verkregen worden met minder gegevens. Redundantie verwijst dus naar het tegenovergestelde van

ruis. Uitbijters zijn monsters met een sterk afwijkende samenstelling. Dit kan veroorzaakt zijn door een inadequate bemonstering of het kan gaan om monsters die een deel van een gradiënt beslaan die in de dataset onderbemonsterd is. Veel multivariate methoden geven in dergelijke omstandigheden weinig bevredigende resultaten en het is daarom belangrijk deze uitbijters te identificeren (zie verder datareductie).

De gegevens kunnen aan de hand van een editor in de computer worden ingevoerd. De manier van werken is uiteraard zeer computer gebonden. Toch moeten de data in een welbepaalde vorm gebracht worden afhankelijk van het programma dat we willen gebruiken.

VOLLE MATRIX VERSUS GECONDENSEERDE MATRIX

De gegevens van een onderzoek kunnen samengebracht worden in een tabel waar per monster de aantallen van elke soort zijn opgenomen. Dit noemen we een volle matrix (Fig. 3.1). Deze matrix kan zowel ingevoerd worden met de soort of met het monster als rij. In dit laatste geval spreekt men van een "transposed matrix". Bij de meeste datasets is het evenwel zo dat we in totaal heel wat soorten aantreffen maar dat er per monster slechts enkele voorkomen maw de totale datamatrix bestaat voor een groot deel uit nullen (Fig. 3.1).

BASISTABEL MET DE DATA

monster	1	2	3	4	5	....
soort 1	1	5	3	13	4	
soort 2	0	0	1	0	0	
soort 3	0	0	0	0	0	
soort 4	1	2	0	9	0	
soort 5	12	1	0	0	0	
soort 6	0	5	0	0	0	

VOLLE MATRIX

```
1  1  0  0  1 12  0
2  5  0  0  2  1  5
3  3  1  0  0  0  0
4 13  0  0  9  0  0
5  4  0  0  0  0  0
:  :
:  i- aantal soort 1
:     t.e.m. soort 6
:
i---- monsternummer
```

CONDENS MATRIX

```
1  1  1  4  1  5 12
2  1  5  4  2  5  1  6  5
3  1  3  2  1        :  :
4  1 13  4  9        :  :
5  1  4              :  :
:  :  :              :  :
:  :  i- aantal -----:--i
:  i---- soortnummer-i
:
i-------- monsternummer
```

Fig. 3.1. Voorbeeld van een datamatrix en de resulterende volle en gecondenseerde matrix. De invoer is op basis van de monsters, de zogenoemde "Transposed form".

Bij een vegetatiekundige studie van bossen in Binnen-Vlaanderen vond Hermy (1985) in een data-matrix van 640 opnames en 328 species 93% nullen. Zo'n volle matrix invoeren is tijdrovend en de opslag neemt heel wat computerruimte in beslag. Daarom kunnen we de matrix, voor stockage, beter omzetten in een gecondenseerde matrix (Fig. 3.1).. Hierin zijn per monster de van nul verschillende waarnemingen opgenomen en wel zo dat we enkel het soortnummer en het aantal opnemen. Hermy (1985), gebaseerd op de reeds vermelde dataset, vond dat een condensed file slechts zo'n 15% van de ruimte van een volle matrix inneemt.

FREEFIELD VERSUS FIXED FORMAT

Zowel een volle als een gecondenseerde matrix kunnen nu nog op verschillende manieren worden ingevoerd zoals is weergegeven in Fig. 3.2. In een freefield invoer zijn alle getallen gewoon gescheiden door een komma of een spatie, in fixed format neemt elke variabele (monster- en/of soortnummer, aantallen) steeds hetzelfde aantal kolommen in beslag. Een fixed format file is veel overzichtelijker om te lezen maar is veel moeilijker in te voeren en neemt meer plaats in dan een freefield matrix.

freefield condens	fixed condens
1,1,1,4,1,5,12	1 1 1 4 1 5 12
2,1,5,4,2,5,1,6,5	2 1 5 4 2 5 1 6 5
3,1,3,2,1	3 1 3 2 1
4,1,13,4,9	4 1 13 4 9
5,1,4	5 1 4

freefield full	fixed full
1,1,0,0,1,12,0	1 1 0 0 1 12 0
2,5,0,0,2,1,5	2 5 0 0 2 1 5
3,3,1,0,0,0,0	3 3 1 0 0 0 0
4,13,0,0,9,0,0	4 13 0 0 9 0 0
5,4,0,0,0,0,0	5 4 0 0 0 0 0

Fig. 3.2. Voorbeeld van de verschillende manieren waarop de data in de computer kunnen worden ingevoerd. De matrices zijn analoog aan die van Fig. 3.1.

Het zal duidelijk zijn dat het van groot belang is om een efficiente soortsnummering bij te houden. Immers wanneer je voor elke nieuwe dataset van een bepaald onderzoek de soorten opnieuw nummert kan je gegevens van veschillende onderzoeken nadien niet samenbrengen en onstaat bovendien op de kortste keren enorme verwarring over de soortsnummers. Het is daarom aangeraden voor elk type onderzoek (bv. vegetatiekunde, macrozoöbenthos etc.) 1 lijst met soortnummers aan te maken wordt die steeds gebruikt wordt.

Voor de verwerkingsprogramma's van de Cornell University (DECORANA en TWINSPAN) wordt een condensed format bestand als invoer gevraagd. Dit bestand (Fig. 3.3) is echter niet freefield: op de eerste lijn komt het aantal soorten, het aantal monsters, de titellijn, het typeformaat van de matrix en de afkorting van de titellijn. Op de tweede lijn komt het inputformat, waarmee de eigenlijke matrix moet ingelezen worden, en het aantal soortsnummer - dichtheid koppels per lijn. Dan volgt de fixed format data-matrix met onderaan de file de soort- en de monsterlabels, elk 8 karakters lang.

```
                                                        :--kolom 71
                                                        :
    6    5VOORBEELD DATAMATRIX IN CORNELL FORMAT ... TCDEMO**
(I2,5(I3,F3.0))
 1  1  1  4  1  5 12
 2  1  5  4  2  5  1  6  5
 3  1  3  2  1
 4  1 13  4  9
 5  1  4
00
SOORT  1SOORT  2SOORT  3SOORT  4SOORT  5SOORT  6
MONSTER1MONSTER2MONSTER3MONSTER4MONSTER5
```

Fig. 3.3 Voorbeeld van een Cornell condense fixed format bestand.

Het is belangrijk om op te merken dat in een Cornell condens fixed format bestand zowel de soorts- als de monsternummers vervangen worden door sequentiële getallen startend vanaf 1. Gezien de soortsnamen en de monsternummers of -labels onderaan de file bijgevoegd krijgen we hier geen interpretatie problemen.

Voor het omzetten van freefield condensed format files in een volle matrix (bijvoorbeeld bruikbaar in SYSTAT of SPSS) kan het programma FULL worden aangewend. Deze volle matrix file kan dan worden omgezet in een condensed CE (Cornell Ecology) format file m.b.v. het programma CONDENS. (of op PC met behulp van het programma COMPOSE (Mohler, 1987).

FORTRAN FORMATING

In een freefield matrix zijn alle waarden gescheiden door een spatie of een komma (een delimiter) en de computer leest van delimiter tot delimiter en geeft de gelezen waarde aan de overeenkomstige variabele. Bij een fixed format matrix is dit iets ingewikkelder. Gezien alle waarden steeds op een vaste plaats (kolom) staan moeten we de computer informatie verstrekken waar hij de waarde van de desbetreffende variabele kan aantreffen. Dit gebeurt conform de fortran programeertaal. Het komt hierop neer dat we de computer vertellen welk type variabele het is en op welke kolommen en lijn de waarde van die variabele te vinden is. We maken vooreerst onderscheid tussen twee types variabelen: reële en gehele getallen (reals en integers). Die worden met F (van Floating point) en I (van Integer) aangegeven. Daarnaast moeten we van elk getal de lengte (in het aantal kolommen) aangeven en voor reals ook het aantal cijfers dat na de

komma staat. Zo I8 een geheel getal van 8 karakters lang en F5.2 een reeël getal van 5 karakters lang waarvan 2 karakters na de komma staan (zoals bv 12.12; let op, de komma wordt dus meegeteld bij het bepalen van de lengte van de variabele). De cijfers na de F of I slaan dus op de lengte van de variabele. Cijfers voor de F of de I geven aan hoeveel variabelen diezelfde lengte hebben. Zo staat 2F5.2 voor '12.1 13.21'. De eerste variabele beslaat dus de eerste 5 kolommen, de tweede de volgende vijf. 12.1213.21 zal met 2F5.2 corect gelezen worden als 12.12 voor variabele 1 en 13.21 voor variabele 2. Wanneer niet alle variabelen van 1 record (een monster bv.) op 1 lijn kunnen dan doen we gewoon verder op de volgende lijn. Dit wordt in het format aangegeven met een / (slash). Kolommen overslaan kan met een X. Zo wil 5X zeggen dat we 5 posities verder springen. Deze afspraken zijn samengebracht in het volgende format (I5,5F5.2/5X,5F5.2) waarmee we de volgende matrix kunnen inlezen:

```
45 1.2 10.12 1.2315.23 1.10
  11.12 1.3010.1  1.15 1.56
```

Merk op dat we voor het einde van het record geen / moeten ingeven. Op het einde van een record gaat Fortran automatisch over naar de volgende regel. Wanneer we met zeer veel variabelen per record te maken hebben dan kan het format zeer lang worden. Wanneer een bepaald patroon op meerder lijnen herhaald wordt, kunnen we dit als volgt aangegeven: (I5,5(5F3.0/),5F3.0) i.p.v. (I5,5F3.0/5F3.0/5F3.0/5F3.0/5F3.0/5F3.0). Merk op dat we de laatste 5 variablen apart vermelden in het format. Immers hadden we 6(5F3.0/) ingegeven dan zou de computer na het inlezen van de laatste waarde weer naar een nieuwe regel springen. Gezien dit bij het einde van een record reeds automatisch gebeurt zouden wij bijgevolg een regel te ver zitten.

4) SCHEMA VAN DE GEGEVENSVERWERKING

Hier willen we een kort schema geven van de verschillende stappen tijdens de verwerking en hoe we van het ene programma naar het andere kunnen overstappen. Gezien er in de loop van de analyses vaak veel files aangemaakt worden is het nuttig om bij de benoeming ervan een vast systeem te gebruiken. We doen hier een suggestie. Bij diverse instellingen (Rijkswaterstaat versus Rijksuniversiteit Gent en Instituut voor Natuurbehoud) zijn verschillende systemen in voege voor het opslaan van de data. Beide systemen worden hier kort besproken. Voor een gedetailleerde beschrijving van enkele specifieke programma's verwijzen we naar appendix 2, een schema van de verwerking is weergegeven in Fig. 4.1.

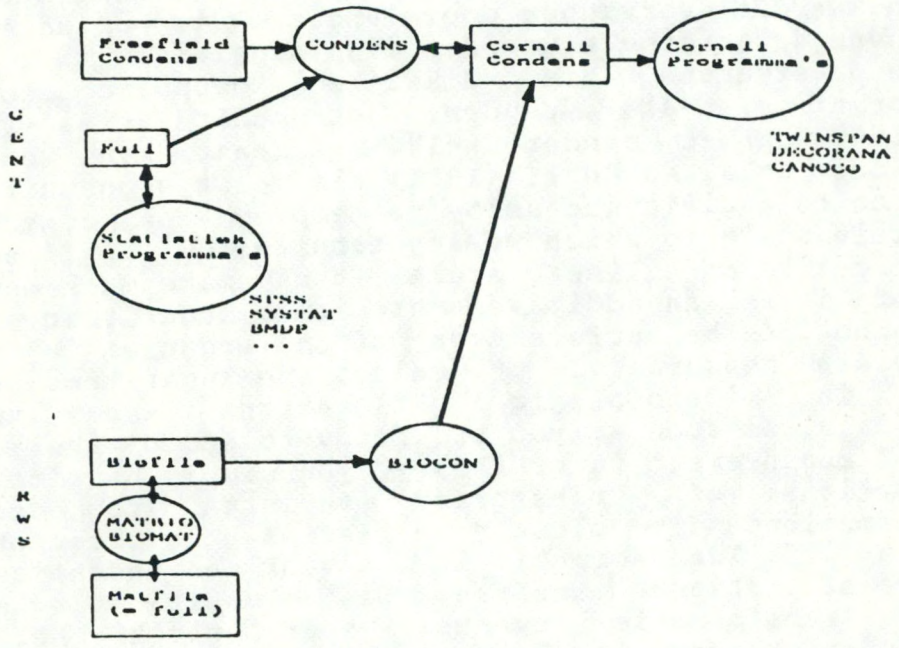


Fig. 4.1 Schema van de diverse programma's en files bij de verwerking van de data. Een cirkel is een programma, een rechthoek een file.

Een matrix wordt ingevoerd in een freefield condens file die via het programma CONDENS kan omgezet worden in een Cornell condens file, een vorm geschikt voor de programma's TWINSPAN, DECORANA en CANOCO. Via Condens kan ook een volle matrix gemaakt worden die dan geschikt is voor diverse statistiek paketten. Een volle matrix kan op zich ook in CONDENS ingelezen worden en op die manier in een Cornell condens file omgezet worden.

De biofiles (van RWS) kunnen via BIOCON worden omgezet naar een Cornell condens file. Via MATBIO kan een mat file gemaakt worden, wat een volle matrix is. Die file kan dan gebruikt worden voor diverse statistiek programmas. De beperking is evenwel dat er maximum 20 soorten kunnen opgenomen worden. Ook via CONDENS kan van de BIOCON file een volle matrix gemaakt worden die voor verdere verwerking geschikt is. Voor de andere programma's die met BIO of MAT files werken verwijzen we naar appendix 2.

5) AANPASSEN VAN DE DATAMATRIX

TRANSFORMATIE

De verdeling van de aantallen per soort in de monsters zijn als regel nagenoeg nooit normaal verdeeld. Nochtans is dit voor vele statistische methoden (met name in parametrische statistiek) vereist. Bovendien kunnen de dichtheden van verschillende soorten ook sterk verschillen. Mochten we bv. in 1 analyse gegevens verwerken van zowel macro- als meiofauna dan hebben we te maken met dichtheidsverschillen van meerdere ordes van grootte. Dit kan de analyse sterk beinvloeden op een manier die niet wenselijk is. Om dit te vermijden kunnen we de data transformeren, d.i. de gemeten waarden vervangen door andere, die op de één of andere manier van de originele waarden zijn afgeleid.

De bezwaren die dikwijls geopperd worden tegen transformaties, als zouden het manipulaties van de data betreffen, waarmee eender welk resultaat kan worden bekomen, worden door Sokal en Rohlf (1981) als volgt weggewuifd :"there is really no scientific necessity to employ the common linear or arithmetic scale to which we are accustomed. ... If a relation is multiplicative on a linear scale, it may make much more sense to think of it as an additive system on a logaritmic scale. ...The square root of the surface area of an organism is often a more appropriate measure.... pH values are logarithms, and dilution series in microbiological titrations are expressed as reciprocals. As soon as you are ready to accept the idea that the scale of measurement is arbitrary, you simply have to look at the distributions of transformed variates to decide which transformations most closely satisfies the assumption of the analysis ...". Transformatie is dus vaak een essentieel onderdeel van elke statistische analyse. Bij classificatie en ordinatie kunnen transformatie evenwel zeer sterk het resultaat beinvloeden. Daarom is het essentieel om het effect van de verschillende transformaties te begrijpen zodat we afhankelijk van de aard van de data kunnen kiezen of we gewild een bepaald effect precies willen vermijden of accentueren. De keuze van een transformatie wordt dus bepaald door 1) de aard van de data en 2) de doelstellingen van de onderzoeker.

In wat volgt geven we een kort overzicht van enkele veel gebruikte transformaties. Voor meer details verwijzen we naar Sokal en Rohlf (1981) en (andere referenties bv. Noy-meir, 1973)

Vele programma's geven de mogelijkheid om transformaties uit te voeren. Die zijn samengevat in Tabel 5.1. Is een transformatie aanwezig in een verwerkingsprogramma zoals DCA dan moet de datamatrix niet op voorhand aangepast worden. Is dit niet zo dan moeten we eerst met een programma de transformatie uitvoeren en daarna de datafile in het goede format brengen voor het toepassingsprogramma dat we willen gebruiken. Voor veel tranformaties kan het programma CONDENS gebruikt worden. De transformatie mogelijkheden in TWINSPAN, DCA en CANOCO zijn beperkt (zie verder onder DCA). Het effect van verschillende transformaties zijn weergegeven in Fig. 5.1.

transformatie:	1	2	3	4	5	6	7
programma							
BIOM	X	X	X	X			X
SPSS	X	X	X				
SYSTAT	X	X		X			
CONDENS	X	X		X	X	X	X

Tabel 5.1. Overzicht van de beschikbaarheid van diverse transformatie mogelijkheden in de in dit rapport vermelde computer programma's. 1: Log of LN (x+1); 2: SQRT (x+1); 3: Box-Cox transformatie; 4: Arcsinus (x); 5: standardiseren (op maximum of op totaal); 6: relativeren (op maximum of op totaal) 7:

STANDARDISEREN EN RELATIVEREN

Wanneer we in de datamatrix grote verschillen hebben in de grootte-ordes van de waarden (bv. voor sommige soorten variëren de aantallen tussen 0 en 10 en voor andere tussen 0 en 10000) of wanneer de aantallen tussen de verschillende soorten onderling niet echt vergelijkbaar zijn (zoals bv. bij bodemvallen) is het aangewezen om de gegevens te standardiseren of te relativeren. Dit is ze uit te drukken als afwijking van het gemiddelde (centreren), het totaal, de maximum waarde of de range (maximum-minimum). Dit kunnen we als volgt voorstellen:

$$X_{ij}' = X_{ij} - \text{Tot}n/n$$

of

$$X_{ij}' = X_{ij} * 100 / \text{Tot}_n (X_{in})$$

of

$$X_{ij}' = X_{ij} * 100 / \text{Max}_n (X_{in})$$

Met X_{ij} : oorspronkelijke waarneming (meting, dichtheid); X_{ij}' : getransformeerde waarde; Totn : sommatie van X_{ij} over n waarden; Maxn : maximum van X_{ij} over n waarden en n is het aantal soorten bij standardisatie of het aantal monsters bij relativering.

In het ene monster bvb. kan het totaal aantal individuen, gesommeerd over alle soorten, 2000 individuen bedragen terwijl in andere monsters slechts 10 of zelfs geen individuen aangetroffen worden. Eén bepaalde soort, die nochtans zeer geregeld werd aangetroffen, werd misschien slechts met maximale dichtheden van 10 individuen per monster gevonden, terwijl andere soorten veel hogere dichtheden bereikten. Door te standardiseren of te relativeren zetten we alle aantallen om in percentages, die voor alle monsters, resp. voor alle soorten, variëren tussen 0 en 100% zodat aan elk monster (resp. aan elke soort) evenveel belang gehecht wordt bij de analyse m.b.v. numerieke technieken, die

anders al teveel beïnvloed zijn door extreme waarden.

Als we monster per monster transformeren, en dus aan elk monster een gelijk gewicht geven onafgezien van de bereikte dichtheden, spreken we van "standardiseren". Elke dichtheid wordt dan uitgedrukt als percentage van het totaal of het maximum binnen dat monster. Een voorbeeld is weergegeven in Tabel 5.2.

A DATAMATRIX

	monster			
soort	1	2	3	4
1	2000	1000	500	0
2	20	10	5	0
3	0	5	0	0
4	2000	0	0	10

B GESTANDARDISEERDE DATAMATRIX

	monster			
soort	1	2	3	4
1	0.48	0.99	0.99	0
2	0.04	0.01	0.01	0
3	0	0.00	0	0
4	0.48	0	0	1.00
SOM	1	1	1	1

Tabel 5.2. Voorbeeld van het standardiseren van een datamatrix. De waarden uit de datamatrix werden gestandardiseerd als het percentage van het monstertotaal (naar Clifford en Stephenson, 1975).

De zeer hoge waarden in monster 1 zijn afgevlakt wat wenselijk is maar aan de andere kant is het verschil tussen monster 2 en 3 zogoed als weggevallen. Het belangrijkste verschil is de grote waarde die aan soort 4 in monster 4 gegeven wordt. Het (ongewenst) teveel gewicht toekennen aan waarden uit soortenarme stations is dan ook het grootste bezwaar tegen standardisatie.

Een voor de hand liggende oplossing voor dit probleem is de waarden vervangen door het percentage op het soortstotaal. Dit transformeren per soort wordt "relativeren" genoemd. Een voorbeeld toegepast op dezelfde datamatrix is weergegeven in Tabel 5.3.

Met deze techniek worden minder abundante soorten even belangrijk gesteld als dominante. Zo krijgt de zeldzame soort 3 nu veel gewicht. Relativeren wordt normaal minder frequent gebruikt bij normale analyse (d.i. klassificatie van de monsters), maar opent interessante perspectieven bij inverse analyse (d.i. van de soorten). Deze transformatie houdt echter het risico in dat monsters waarin zeer zeldzame soorten voorkomen (niet lage aantallen, maar lage frequentie) sterk onderscheiden worden van de andere (precies omdat een lage frequentie vervangen wordt door een hoog percentage). Daarom wordt in een aparte paragraaf besproken hoe zeldzame soorten of slecht geprofileerde monsters te elimineren.

A DATAMATRIX

	monster			
soort	1	2	3	4
1	2000	1000	500	0
2	20	10	5	0
3	0	5	0	0
4	2000	0	0	10

B GERELATIVEERDE DATAMATRIX

	monster				
soort	1	2	3	4	SOM
1	0.57	0.29	0.14	0	1
2	0.57	0.29	0.14	0	1
3	0	1.00	0	0	1
4	0.99	0	0	0.01	1

Tabel 5.3. Voorbeeld van het relativeren van een datamatrix. De waarden uit de datamatrix werden gerelativeerd als het percentage van het soortstotaal (naar Clifford en Stephenson, 1975).

Eerst relativeren op het soortsmaximum en vervolgens standaardiseren op het monstertotaal (de zogenaamde dubbele of Bray Curtis transformatie) (Cottam et al. ,1978) wordt door Faith et al.(1987) als een superieure standardisatiemethode beschouwd.

In een aantal bijzondere gevallen is het niet alleen raadzaam, maar zelfs noodzakelijk om te relativeren of te standardiseren. Bij braakballenonderzoek bijvoorbeeld kan een braakbal niet als een gestandardiseerd monster beschouwd worden. Een logische oplossing voor dit probleem vormt het samennemen van alle braakballen van een bepaalde vindplaats en het uitdrukken van het aantal aangetroffen individuen van een soort als percentage op het totaal aantal individuen op die vindplaats (standardiseren). Een ander voorbeeld waarbij gerelativeerd moet worden, vormt het pedofaunaonderzoek, waar organismen met verschillende activiteitspatronen - en dus met een verschillende kans om gevangen te worden - bemonsterd worden met bodemvallen. Als we dergelijke data willen gebruiken bij een vergelijking van verschillende stations over een zelfde bemonsteringsperiode (waarin de verschillende soorten dus ongeveer eenzelfde activiteitspatroon vertonen in de verschillende stations), moeten we de verschillen tussen de soorten onderling wegwerken d.m.v. relativeren. Het blijft in veel gevallen evenwel een zeer arbitraire zaak of we al of niet moeten standardiseren of relativeren.

Z-SCORES

In een volgende stap kan men de afwijking van de gemeten waarde uitdrukken in verhouding tot de totale variatie tussen de

variabelen ("standardiseren tot gemiddelde 0 en variantie 1"). De resulterende waarden worden ook wel Z-scores of standaard-scores genoemd. Enige begripsverwarring met standardizeren is hier mogelijk. Ze zijn bekomen door de gemiddelde waarde af te trekken van de gevonden waarde en daarna te delen door de variantie. Dit is een veel gebruikte techniek als het variabelen betreft met bijvoorbeeld andere meeteenheden. De resulterende distributie is normaal verdeeld en deze transformatie laat dan ook het gebruik van parametrische tests toe. Bovendien kunnen de waardes van de verschillende variabelen nu onderling vergeleken worden.

$$X_{ij}' = (X_{ij} - X_{in}/n) / S_i$$

of

$$X_{ij}' = (X_{ij} - Min_n) / abs(Max_n - Min_n)$$

(transformatie volgens Gower, 1971)

Met S_i : Standaardafwijking van X_{ij} over n waarden
Min_n : Minimum van X_{ij} over n waarden

LOGARITMEREN VAN DE DATA

In ecologisch onderzoek vertonen de meeste variabelen vaak "skewed" (scheve) distributies en is de variantie niet onafhankelijk van het gemiddelde. "Whenever the mean is positively correlated with the variance, the logaritmic transformation is quite likely to remedy the situation and make the variance independent of the mean. Frequency distributions skewed to the right are often made more symmetrical by logaritmation" (Sokal en Rohlf 1981). Logaritmeren van de data is dan ook een veel gebruikte transformatie methode, die vaak ook toelaat om te voldoen aan bepaalde voorwaarden van statistische tests (ANOVA,..). Gezien de logaritme van 0 niet gedefiniëerd is, en er wel nullen aanwezig zijn in de dataset, nemen we de logaritme van (x+1). De keuze van het getal hangt evenwel af van de data zelf. Wanneer we allemaal zeer kleine getallen hebben (bv. 0.001 tot 2 bv.), dan is het beter om een zeer klein getal te kiezen, hebben we te doen met veel grotere getallen dan kunnen we beter een groter getal nemen. Het is evenwel zo dat wanneer er veel nullen aanwezig zijn in de dataset we op basis van logaritmeren nog steeds geen normale verdeling bekomen.

$$X_{ij}' = \log (X_{ij} + 1)$$

of

$$X_{ij}' = \ln (X_{ij} + 1)$$

VIERKANTSWORTEL TRANSFORMATIE

Tellingen, zoals bvb. het aantal insecten op een blad of het aantal bloedcellen in een hematocytometer, zijn dikwijls verdeeld volgens de Poisson distributie (variantie = gemiddelde). In dat geval is de vierkantswortel transformatie aan te bevelen om de data te normalizeren. In verband met nulwaarnemingen geldt dezelfde opmerking als bij logaritmeren.

$$X_{ij}' = \sqrt{X_{ij}}$$

In veel gevallen blijkt een vierdemachtswortel een veel beter resultaat op te leveren dan een vierkantswortel. Nog diverse andere transformaties op basis van een machtsfunctie zijn denkbaar. Door Downing (1979) werd voorgesteld om de b waarde van de Taylor Power Law te gebruiken. Dit verbetert de resultaten echter niet merkbaar (Dereu en Meire, 1987).

BOX-COX TRANSFORMATIE

Vaak is er niet een a priori reden om de een of de andere transformatie toe te passen. De Box-Cox transformatie is nu precies een procedure om de beste transformatie te schatten die de gegevens naar een normaal verdeling omzet. De transformatie wordt gekozen uit een familie machtsvergelijkingen

$$X_{ij}' = (X_{ij}\uparrow^{lambda} - 1)/lambda \quad (\text{voor } Lambda \neq 0)$$

$$X_{ij}' = \ln X_{ij} \quad (\text{voor } Lambda = 0)$$

en is bijgevolg een algemenere benadering van wat hierboven werd aangehaald.

ARCUS SINUS TRANSFORMATIE

Deze transformatie, ook gekend als de angulaire transformatie is bijzonder geschikt wanneer we te maken hebben met percentages (bv. bedekkingen) en verhoudingen. Deze transformatie vindt $\theta = \arcsin \sqrt{p}$, waar p de verhouding of percentage is. Arcus sinus is synomiem voor de inverse van de sinus, wat staat voor de hoek waarvan de sinus gegeven kwantiteit is. De Arcsin van 0.431 is 41.03 wat de hoek is waarvan de sinus $= \sqrt{0.431}$. Deze transformatie trekt de uiteinden van de verdeling uiteen en duwt het midden samen.

$$X_{ij}' = (\text{Arcsin } (X_{ij}))^{1/2}$$

Deze transformatie wordt veel in de vegetatiekunde toegepast.

BINAIRE TRANSFORMATIE

Deze transformatie is in feite niets meer dan het vervangen van de originele waarneming door een nul of een één, en geeft dus meestal aan of een soort al dan niet aanwezig was in het studiegebied. Deze transformatie is in feite geïncorporeerd in een aantal similariteitsindices (cf.verder).

GROEPEREN IN KLASSEN

Bij het verzamelen van gegevens is het dikwijls gemakkelijker met abundantieklassen of bedekkingsgraden te werken. Ook achteraf is het hergroeperen in klassen dikwijls zinvol.

VERGELIJKING VAN ENKELE TRANSFORMATIE

In Fig. 5.1 is de verdeling van de dichtheden van een soort over alle monsterpunten weergegeven (het betreft de dichtheden van Heteromastus filiformis in 95 monsterpunten in de Westerschelde). De ruwe data (Fig. 5.1a) tonen zeer duidelijk dat de gegevens niet normaal verdeeld zijn (Tabel 5.4). Door het berekenen van Z-scores verandert de verdeling, in tegenstelling tot wat men meestal denkt, helemaal niet. Enkel wordt het gemiddelde 0 en de standaard afwijking 1. Logaritmeren van de data brengt een duideljke verandering teweeg in de verdeling maar het is duidelijk dat de nulwaarden (vervangen door 1) sterk afwijken (Fig. 5.1b) en de verdeling is nog steeds significant afwijkend een normaalverdeling (Tabel 5.4). Vervangen van 0 door 10 levert een verdeling op die wel normaal verdeeld is (Tabel 5.4). De vierkants en vierdemachtswortel en de Box Cox transformatie (met Lambda = 0.18313) brengen een steeds betere benadering van de normaal verdeling met zich mee (Tabel 5.4 en Fig. 5.1). Het nadeel van de Box Cox transformatie is dat de transformatie waarde voor elke soort veschilt. Wanneer we op deze manier alle soorten willen transformeren is dit zeer arbeidsintensief.

transformatie	K-S waarde	Probabiliteit
ruwe data	2.581	< 0.001
Z-scores	2.581	< 0.001
Log (x + 1)	2.901	< 0.001
Log (x + 10)	0.787	NS
Vierkantswortel	1.525	< 0.05
Vierdemachtswortel	0.994	NS
Box Cox (= 0.18)	0.828	NS

Tabel 5.4. Overzicht van de resultaten van een Kolmogorov Smirnof test voor de vergelijking van de data met een normaal verdeling. De testwaarde en de bijhorende probabiliteit zijn weergegeven. Slechts in drie gevallen is de getransformeerde verdeling niet statistisch te onderscheiden van een normaal verdeling.

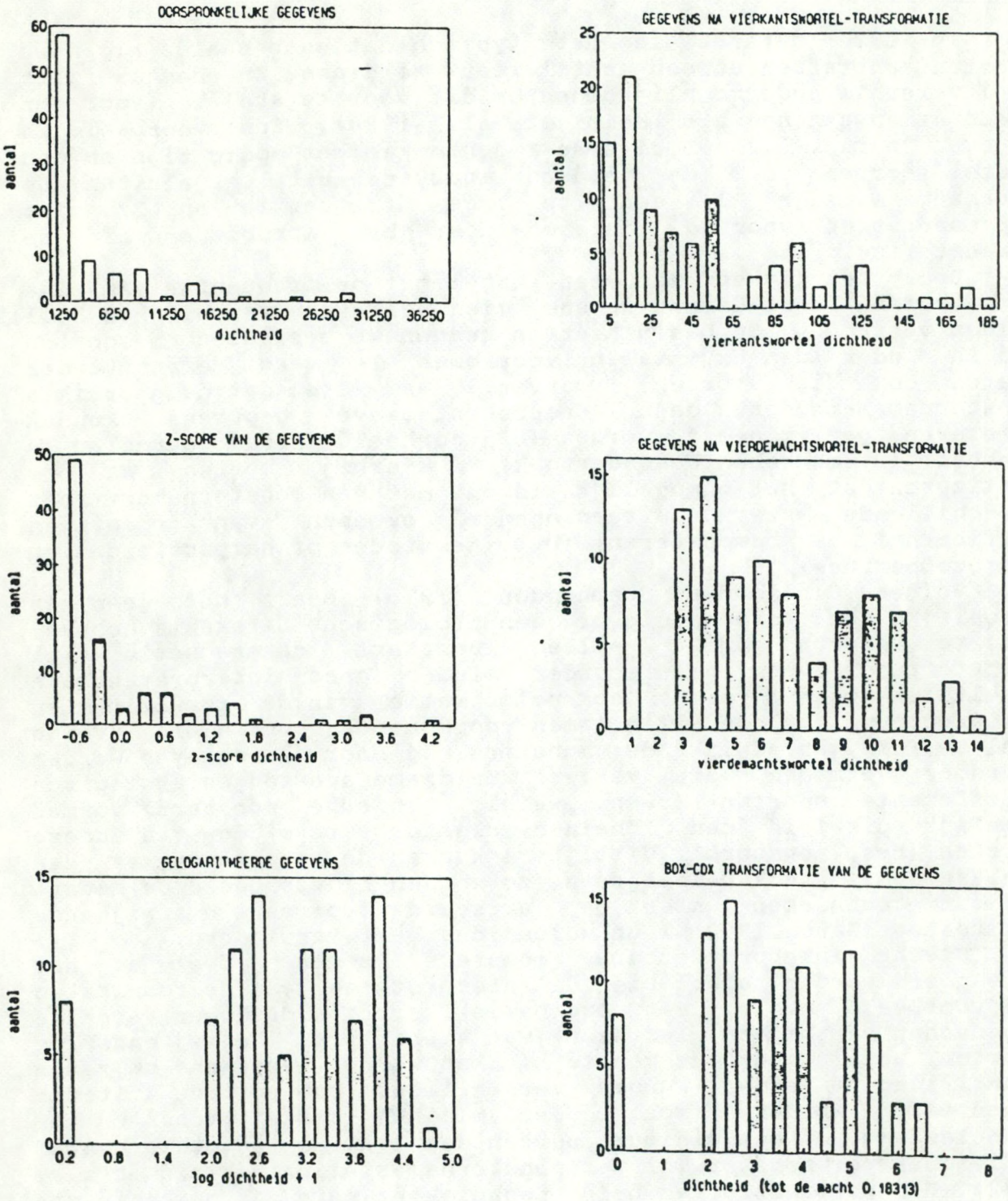


Fig. 5.1 Overzicht van het effect van verschillende transformaties op een dataset. a ruwe data; b log (x+1); c Vierkantswortel; d Vierdemachtswortel; e z-scores; f Box-Cox transformatie.

DATA REDUCTIE

In elke dataset is het typisch dat we bepaalde algemene soorten aantreffen en een aantal zeer zeldzame. Er bestaan zeer veel verschillende mogelijkheden om dit voor te stellen (voor een goede samenvatting zie Heip et al. 1988). Een voorbeeld is gegeven in Fig. 5.2. Om diverse redenen kan het nodig zijn om een aantal soorten voor de verdere analyses uit te sluiten. De voornaamste zijn 1) een te lage frequentie en 2) niet representatief voor de dataset door bv. problemen bij de bemonstering.

Door het maken van een tabel met de frequentie van elke aangetroffen soort kunnen we snel zien welke soorten slechts heel zelden voorkomen. Op basis hiervan kunnen we beslissen om soorten die in minder dan x stalen voorkomen te verwijderen. Verder moeten ook die soorten, waarvan we weten dat de gebruikte monsternamemethodes geen representatieve gegevens kunnen opleveren, geëlimineerd worden (bijvoorbeeld vissen, garnalen en krabben in macrobenthosonderzoek). Hierbij moeten we ons realizeren dat het onmogelijk is om met één monsternamemethode verschillende sterk uiteenlopende groepen van organismen (efficiënt) te bemonsteren (bvb. nematoden of harpacticiden en macrozoöbenthos).

Volgens Clifford en Stephenson (1975) heeft het geen zin dergelijke data, die weinig of geen biologische betekenis hebben, uit te werken. Niet alleen bespaart datareductie veel computertijd, en dat zonder minder goed interpreteerbare resultaten op te leveren, ook polarisatie van de resultaten en het groeperen van toevallig samen voorkomende soorten bij inverse analyse, of van afwijkende monsters bij normale analyse worden hierdoor vermeden (zie verder). Zeldzame soorten en ekologisch indifferente soorten (resp. weinig typische monsters) vormen namelijk dikwijls een "chained cluster" te midden van andere soorten (resp.monsters). Uitbijters kunnen de resultaten van een analyse soms sterk beïnvloeden (zo worden bijvoorbeeld de meeste ordinatietechnieken sterk gepolariseerd door enkele afwijkende entiteiten (Gauch,1982)) en worden dus best verwijderd.

Slecht interpreteerbare monsters kunnen ofwel ad hoc opgespoord worden, ofwel bij de interpretatie van de resultaten (bijvoorbeeld m.b.v. een ordinatie). Precies de hoger vermelde eigenschap van bijeen clusteren van uitbijters (hetzij onder de soorten, hetzij onder de monsters), kan worden aangewend om ze te identificeren. Verder kunnen verschillende objectieve kriteria gehanteerd worden, zoals een minimum aantal vondsten of vindplaatsen, of een minimum percentage van het totale aantal individuen per monster (of per monsternamestation) (Field et al., 1982). Een combinatie van beide technieken (namelijk op basis van een dendrogram en op basis van een arbitrair cutlevel) kan voorkomen dat zeldzame, maar ekologisch significante soorten geëlimineerd worden.

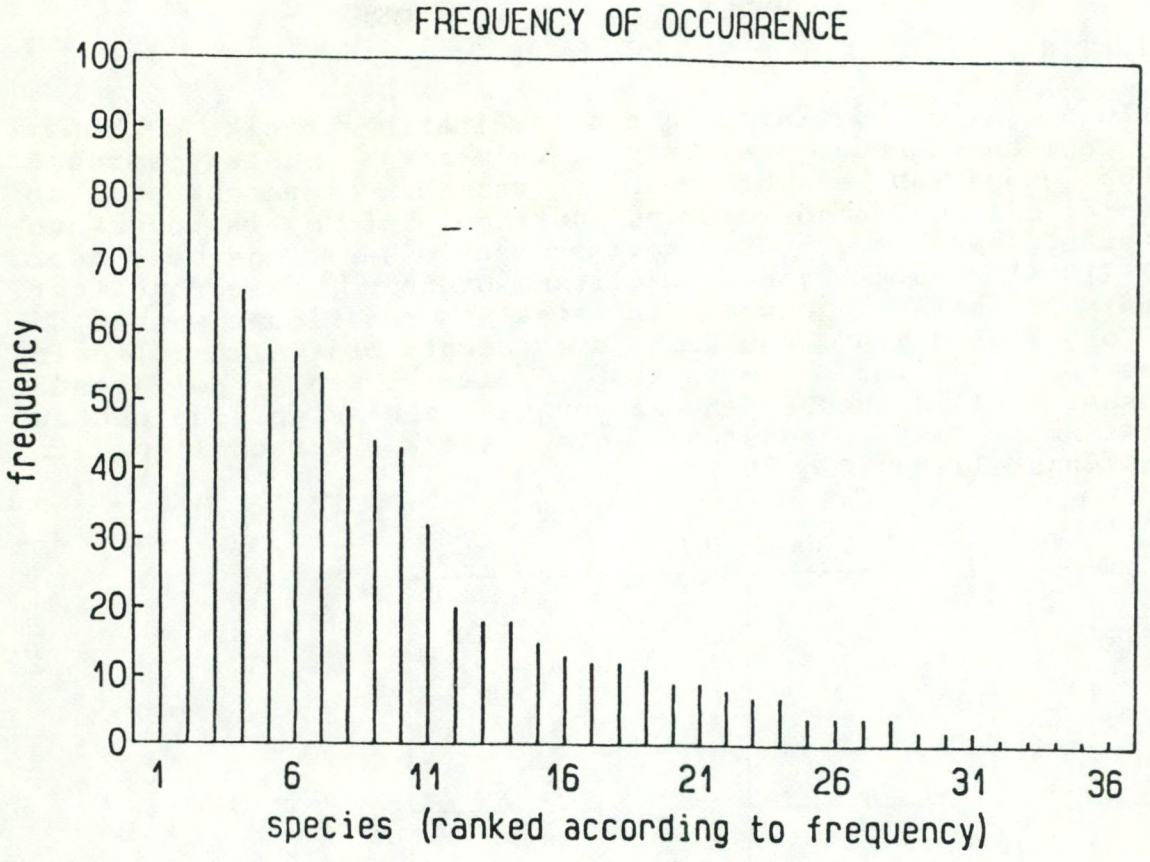


Fig 5.2. Frequentie van voorkomen van diverse soorten. De soorten werden gerankt en voor elke soort wordt de frequentie van voorkomen in de dataset weergegeven.

6) ORDINATIE

Orloci (1973) definiëert een ordinatie als elke methode, waarbij gegeven eenheden als abstract ruimtelijke punten geordend worden op grond van één of meerdere van hun eigenschappen, op zo'n manier dat hun rangschikking nuttige informatie over hun verwantschap kan geven. De term ordinatie werd ingevoerd door Goodall (1954) en stamt van het duitse 'Ordnung'. Het resultaat van een ordinatie in twee dimensies is een figuur waarin de punten zo geördend zijn dat zij die dicht bij elkaar liggen overeenstemmen met monsters die een gelijkende soortensamenstelling hebben en dat ver van elkaar gelegen punten overeenstemmen met monsters die sterk verschillen in soortensamenstelling (Fig. 6.1).

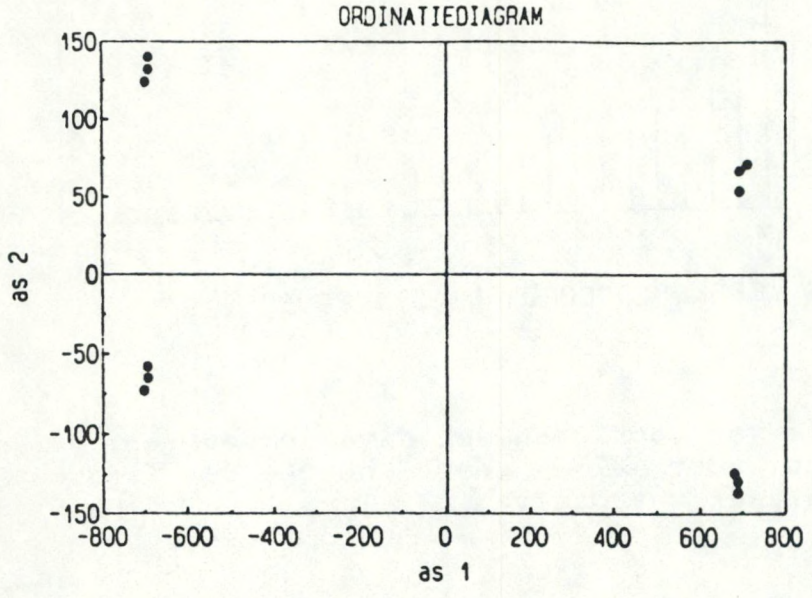


Fig. 01. Hypothetisch voorbeeld van een ordinatie waarin vier groepen van sterk gelijkende monsterpunten zijn weergegeven.

Meer algemeen kunnen monsters (resp.soorten) voorgesteld worden als een puntenwolk in een n-dimensionale ruimte (met n soorten resp. n monsters), waarbij elke soort (resp. monster) één as (dimensie) voorstelt. Konkreet wordt een monster dan gedefinieerd door n coördinaten namelijk de dichtheden (of de biomassa's) voor de n soorten. Deze voorstellingsruimte bevat een bepaalde structuur, die we er omwille van het grote aantal dimensies niet uit kunnen afleiden. Gauch (1982) beschouwt ordinatie als een mathematisch middel dat de dimensionaliteit van een datamatrix zodanig reduceert, dat de structuur eenvoudig kan voorgesteld worden, terwijl toch een minimum aan informatie verloren gaat. Meer algemeen kunnen we het stellen als "The purpose of ordination , beyond arrangements of ecological significance, is that of science: understanding -in this cases understanding of the complex patterns of natural communities in relation to environments" (Austin, 1976).

Klassificaties daarentegen, delen deze entiteiten, op basis van de ekologische afstand (dis-)similariteit (die een één-dimensionele maat is), op in verschillende klassen, waarbij een

deel van de informatie, vervat in de overige dimensies verloren gaat. Deze "klassering" leunt weliswaar dichter aan bij het menselijk denken dan ordinatie (Gauch, 1982), maar is misschien minder geschikt voor het bestuderen van gemeenschappen langs continue gradiënten. Niettemin kan klassificatie, complementair met ordinatie, zeer nuttig zijn om de gedachten te vestigen en te kunnen werken met groepen in plaats van met een moeilijk te hanteren continuüm.

In Fig. 6.2 is weergegeven hoe ordinaties gebruikt worden in ecologisch onderzoek. Ecosystemen zijn zeer complex: ze bestaan uit veel interagerende biotische en abiotische factoren. Om een idee te krijgen van de structuur in de gemeenschap en welke factoren die structuur bepalen gaan we meestal op meerdere plaatsen monsteren. We verzamelen gegevens over de biota (aan- of afwezigheid of abundantie van de soorten) en de abiota (metingen van diverse variabelen of een klasseïndeling). Gezien het aantal soorten vaak groot is (> 30) proberen we de data samen te vatten via een ordinatiemethode in een ordinatie diagram. Dit diagram kan dan geïnterpreteerd worden op basis van de beschikbare gegevens over omgevingsfactoren. Dit kan op een informele manier, wanneer geen echte metingen van omgevingsfactoren voorhanden zijn, of op formele manier (bv. op basis van correlatie analyse), wanneer die wel gemeten zijn. Deze twee-staps benadering wordt, naar Whittaker (1967) "INDIRECT GRADIENT ANALYSIS" genoemd.

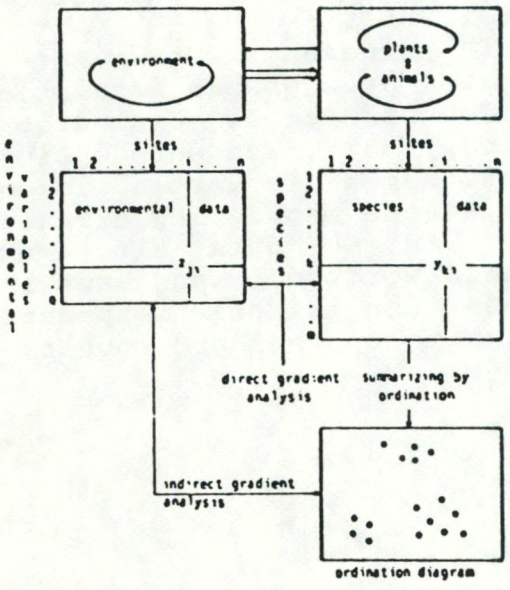


Fig. 6.2. Plaats van ordinatie in "community ecology", met een overzicht van de data en direct en indirect gradient analysis.

Daartegenover staat "DIRECT GRADIENT ANALYSIS". Hiervoor zijn uiteraard gedetailleerde metingen van omgevingsfactoren nodig. Hun invloed op de soorten wordt dan bv. via regressieanalyse onderzocht. Hoewel intuïtief direct gradient analysis veel efficiënter lijkt noemt Ter Braak (1987) drie redenen om indirect gradient analysis te verkiezen. Ten eerste is de soortsamenstelling vaak gemakkelijk te bepalen terwijl het meten van omgevingsfactoren vaak lastig of methodologisch moeilijk is.

Bovendien is de keuze van de te meten omgevingsfactoren meestal subjectief en het voordeel van indirect gradient analysis is dat indien we geen correlaties vinden tussen de ordinatie assen en de gemeten omgevingsfactoren we zeker zijn dat belangrijke factoren niet gemeten werden. Ten tweede kan het voorkomen van individuele soorten te onvoorspelbaar zijn om de relatie met onderliggende factoren te bepalen via directe wegen terwijl precies meer algemene patronen van het samen voorkomen van diverse soorten ons kunnen helpen in het ontdekken van de onderliggende soorts-omgevings relaties. Ten derde kan de interesse vooral gaan naar de combinaties van soorten, zoals in landschapsecologie, en niet zozeer naar het gedrag van individuele soorten. Naar onze mening gaat het eerste argument niet op.

De meest populaire ordinatiemethodes zijn PCA (Principal Component Analysis), CA (Correspondence Analysis) en aanverwante technieken (WA of Weighted Averaging en DCA of Detrended Correspondence Analysis) en MDSCAL (Multidimensional Scaling).

Tussen regressie analyse en ordinatie in staat de techniek van Canonische Ordinaties. Dit zijn ordinaties die omgevormd zijn in multivariate direct gradient anlyses: ze houden simultaan rekening met veel soorten en veel omgevingsfactoren. De hoofdbedoeling hiervan is de belangrijkste patronen in de relaties tussen de soorten en de omgeving te ontdekken. Dit is de Canonical Correspondance analysis en varianten opgenomen in het programma CANOCO van Ter Braak (1986).

Vooraleer we evenwel ingaan op de diverse ordinatietechnieken willen we hier even enkele assumpties die ten grondslag liggen aan de methoden bespreken. Ordinatie kunnen we op twee manieren zien (Prentice, 1977). Ten eerste simpelweg als een manier om multivariate data samen te vatten in een scatter diagram. Meer ambitieus is de tweede benadering die aanneemt dat er een onderliggende structuur in de data zit, maw dat het voorkomen van de soorten bepaald wordt door een aantal onbekende omgevingsfactoren en dit volgens een simpel responsmodel. Deze structuur moet dan door ordinatie opgehelderd worden. Dit wordt geïllustreerd in Fig. 6.3.

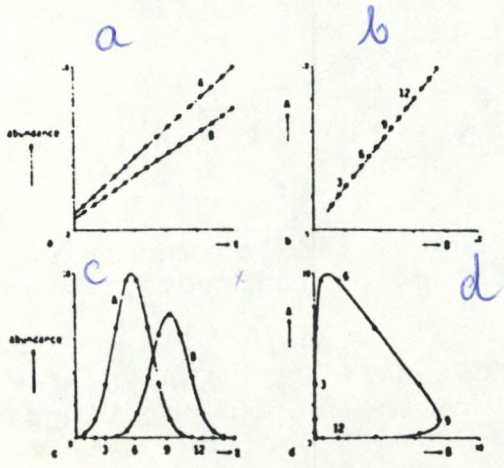


Fig. 6.3. Twee mogelijke responscurves van 2 soorten ten aanzien van een latente variabele x (a, c), en de overeenkomstige plots van de abundantie van soort ten opzichte van soort b (b, d). Het ordinatie probleem is om de patronen in a en c af te leiden uit de data weergeven in b en d.

In deze figuur wordt een voorbeeld gegeven van een lineaire (Fig. 6.3a) en een unimodale responscurve (Fig. 6.3c) van de soorten ten aanzien van een latente variabele x. Het ordinatie probleem is dus om een schatting te maken van deze curves op basis van de informatie samengevat in Fig. 6.3b en d, namelijk de abundanties van de verschillende soorten in de monsters. Het probleem is dus vergelijkbaar met regressie, uitgezonderd dat de onafhankelijke variabelen onbekende omgevingsfactoren zijn. Deze onbekende, latente variabelen worden zo bepaald dat ze de soortsdata best verklaren. PCA is gebaseerd op een lineair model, CA, DCA en CANOCO op een unimodaal responsmodel.

Het bestaan van verschillende methoden brengt ons automatisch op de vraag wat is de "best methode". Hierover werden reeds heel wat testen gepubliceerd. Eerst zullen we de diverse methoden situeren en daarna op de voordelen van de diverse technieken ingaan.

PRINCIPAL COMPONENT ANALYSIS (PCA).

Principal Component Analysis werd in de vegetatiekunde geïntroduceerd door Goodall (1954), hoewel ze reeds veel vroeger op punt gesteld werd door Pearson (1901) en Hotelling (1933, beide in Gauch, 1982). Sinsdien werd deze methode veelvuldig in ecologische studies gebruikt. De nieuwigheid van PCA bestond erin dat het de eerste ordinatietechniek in de ecologie was die enkel en alleen afgeleid werd uit de datamatrix (zonder gewichten of wat dan ook toe te voegen). Het werd beschouwd als een zeer objectieve techniek (Gauch, 1982). De bedoeling en de naam van PCA kunnen we best omschrijven zoals Hotelling (1933, in Gauch, 1982):

> "Als we van een bepaalde entiteit meerdere variabelen (zeg n) meten, dan is de kans groot dat een aantal van die variabelen onderling gecorreleerd zullen zijn. Daarom zoeken we naar een meer fundamentele set variabelen (liefst minder dan n) die de data evengoed beschrijven. Die variabelen (dimensies), die onafhankelijk van elkaar zijn, worden "componenten" genoemd. Sommige componenten kunnen belangrijker zijn dan andere, vandaar de naam "principal component analyse".

De data moeten aan verscheidene voorwaarden voldoen vooraleer PCA mag worden toegepast. Belangrijk is dat de gegevens normaal moeten verdeeld zijn. Het onderliggende model is een lineair responsmodel (zie Fig. 6.3a) en dit heeft belangrijke consequenties, ook al wordt PCA slechts als beschrijvende techniek gebruikt.

PCA is een eigenvectormethode. De afleiding ervan kan op diverse manieren begrijpelijk gemaakt worden. Het PCA algoritme werd geometrisch op een bijzonder duidelijke manier voorgesteld door Gittins (1969) waarvan we hier een samenvatting geven. Voor meer technische uitleg verwijzen we naar Orloci (1978) en Pielou (1984). Een "two-way weighted summation method" algoritme, zoals beschreven door Ter Braak (1987), introduceert PCA op een manier die veel gemeen heeft met CA (zie verder), hij beschrijft PCA ook

als een extensie van lineaire regressie.

GEOMETRISCHE KENMERKEN VAN DE DATASET

De algemene vorm van een data-matrix kunnen we gemakkelijk als volgt voorstellen (Fig. 6.4): de rijen van de matrix geven de variatie aan in de ide soort over de NR monsters, terwijl de kolommen de samenstelling geven van het jde monster in termen van NS soorten.

		monsters				
		1	2	3	....	NR
s	1	$X_{1\,1}$	$X_{1\,2}$		....	$X_{1\,NR}$
o	2	$X_{2\,1}$				$X_{2\,NR}$
o	3					
r	4					
t	.					
e	.					
n	.					
	NS	$X_{NS\,1}$	....			$X_{NS\,NR}$

Fig. 6.4: Voorbeeld van een datamatrix met NR monsters en NS soorten. Soortsscores worden aangegeven door $X_{i\,j}$, de score van de ide soort in de jde opname.

De rijen kunnen naast soorten ook andere ecologische variabelen omvatten. Alhoewel de waargenomen waarden in hun originele vorm kunnen geanalyseerd worden, worden bij PCA de gegevens gewoonlijk voorgesteld als standaard- of z-scores. In een opnameruimte, een multidimensionale ruimte, kan nu iedere soort voorgesteld worden door een punt of een vector (Fig. 6.5).

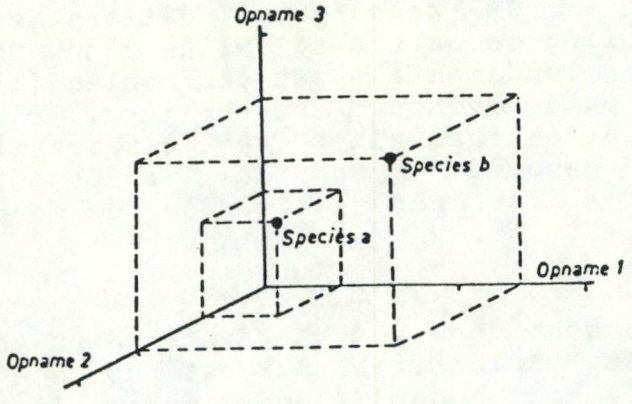


Fig. 6.5. Hypothetisch voorbeeld van een opnameruimte met voorstelling van twee soorten. De coördinatenassen komen overeen met de 3 monsters of opnames en de corresponderende schaal is gewoonlijk een maat voor de abundantie-dominantie (bv. % bedekking of dichtheid). De positie van de punten wordt bepaald door de bedekking van de soorten in iedere opname (naar Gittins, 1969).

De rijvector Z_i : ($Z_{i\,1}$, $Z_{i\,2}$, ... Z_{NS}) is dan de coördinaat van de soort i in deze Euclidische ruimte met NR dimensies. De positie van iedere soort wordt dus slechts bepaald door lineaire

combinatie van de waargenomen waarden over de NR-opnamen. Soorten die wat betreft gedrag sterk op elkaar gelijken bevinden zich in de ruimte ook dicht bij elkaar en omgekeerd. Als ieder punt verbonden wordt met de oorsprong (Fig. 6.6) krijgt men een vectorvoorstelling van de soorten of variabelen.

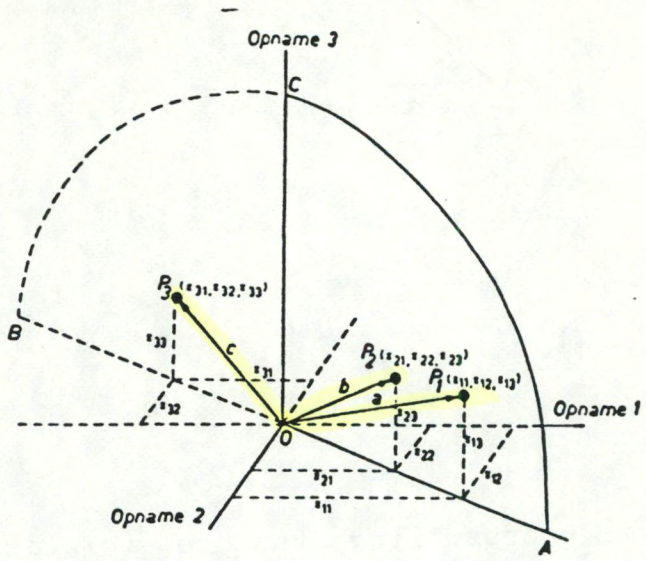


Fig. 6.6. Vectorvoorstelling van driesoorten (P1, P2, P3) in de opnameruimte. De positie van de drie punten wordt bepaald door de abundantie-dominantie van de soorten in de monsters of opnames (naar Gittins, 1969).

In Fig. 6.5 en 6.6 worden slechts enkele opnamen en soorten voorgesteld. Een alternatieve geometrische voorstelling van een data-matrix is mogelijk waarbij de soorten als assen fungeren en de opnamen als punten in de soortsruimte voorgesteld worden.

GEOMETRISCHE STRUCTUUR VAN EEN CORRELATIEMATRIX

De interesse in de vectorvoorstelling van veldwaarnemingen (monsters, soorten) ligt in de ruimtelijke relaties tussen de punten. De relaties kunnen gemeten worden in termen van afstanden tussen de punten of in termen van de hoeken tussen de vectoren (Van Groenewoud, 1965). Er bestaat een eenvoudige relatie tussen de vectorvoorstelling van een paar soorten en de correlatiecoëfficient (Fig. 6.7). Als de variabelen (soorten) uitgedrukt zijn onder de vorm van standaardscores dan is de cosinus van de hoek tussen de twee vectoren in de opnameruimte gelijk aan de waarde van de correlatiecoëfficient tussen de twee variabelen (1).

$$\cos \alpha_{jk} = r_{jk} \qquad (1)$$

met α_{jk} = de hoek tussen de vectoren j en k en r_{jk} = produkt moment correlatiecoëfficient tussen de variabelen j en k over NR monsters of opnames.

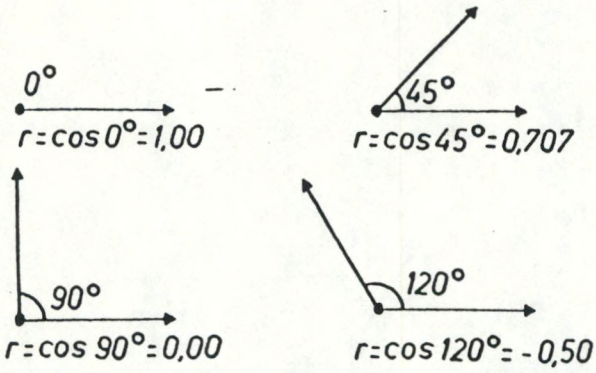


Fig. 6.7. Ruimtelijke voorstelling van de relatie tussen de vectorvoorstelling van de soorten in een opnameruimte en de correlatiecoëfficient tussen de soorten.

De correlatiecoëfficient r kan dus gebruikt worden om de relatie te analyseren tussen de soorten (opgevat als eindpunten in een opnameruimte van vectoren vertrekkende van de oorsprong). Hoe kleiner de hoek tussen de soorten, hoe groter de positieve waarde van de corresponderende correlatiecoëfficiënt (Fig. 6.8).Punten gescheiden van elkaar door een hoek van 90 stellen niet-gecorreleerde taxa voor.

Door uitbreiding van deze vaststelling voor meer dan twee variabelen (NS) vindt men een perfecte overeenkomst tussen de vectorvoorstelling van een data-matrix en de matrix van inter (product-moment) correlaties tussen de soorten. In veel toepassingen worden correlaties tussen soorten veelvuldig vastgesteld. In termen van puntvoorstellingen van de soorten betekent dit dat de punten zich in mindere of meerdere mate bevinden in ellipsoidale clusters (groepen) en niet homogeen verdeeld zijn over de NR-ruimte. In termen van vectoren betekent dit een associatie van vectoren in groepen van pijlen. In Fig. 6.8 en 6.9 wordt dit voorgesteld voor een aantal species in een ruimte van drie, respectievelijk twee opnamen.

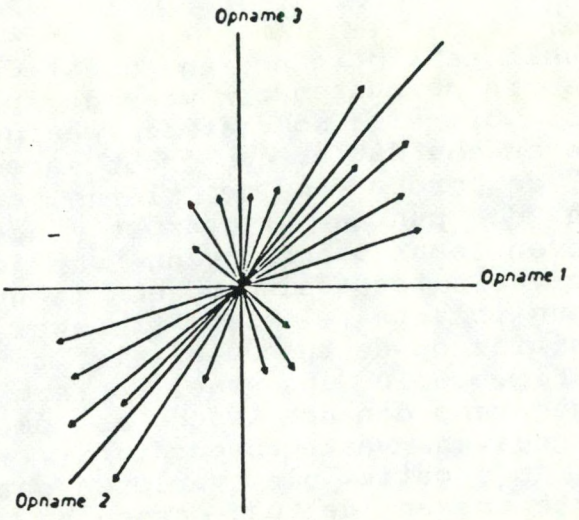


Fig. 6.8. Vectorvoorstelling van taxa in opnameruimte (NR=3). Het bestaan van positieve correlaties wordt aangeduid door hoeken tussen de vectoren (naar Gittins, 1969).

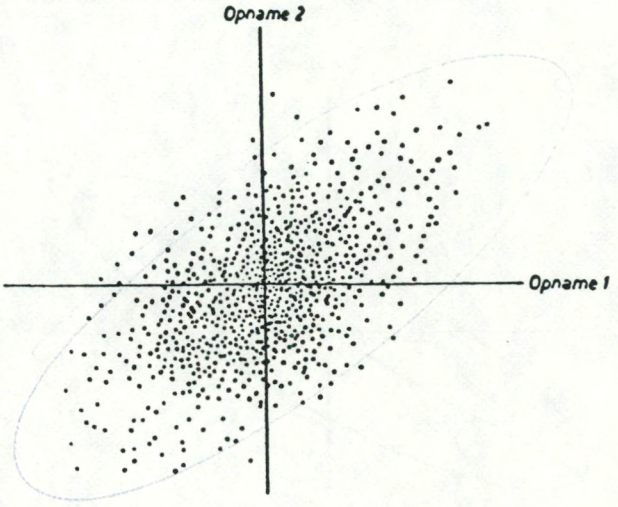


Fig. 6.9. Puntvoorstelling van taxa in een opnameruimte (NR=2). Het bestaan van correlaties tussen veel punten leidt tot een ellipsoidale structuur van de puntenwolk (naar Gittins, 1969).

HOOFDCOMPONENTEN TRANSFORMATIE ('PRINCIPAL AXES')

De berekening van de produkt-moment correlatiematrix tussen soorten is gewoonlijk de eerste stap in PCA. Soms vertrekt men van een similariteitsmatrix tussen opnames; voorwaarde is dat een geschikte coëfficiënt gekozen wordt (zie Orloci, 1966, 1967).

Aanwezigheid van min of meer hoge correlaties in de correlatiematrix houdt in dat de NS-punten (of vectoren) zich in een beperkt deel van de NR-dimensionele ruimte bevinden. Dit maakt het interessant om nieuwe coördinatenassen te construeren doorheen die gedeelten van de opnameruimte waar de punten zich concentreren. De punten kunnen dan efficiënter beschreven worden in relatie tot de nieuwe assen dan in relatie tot de oude assen. PCA is nu in essentie een techniek om een nieuw

coördinatenstelsel te introduceren zodat de nieuwe assen lopen door die plaatsen in de puntenwolk waar de spreiding in de punten het grootst is. Dit is schematisch weergegeven in Fig. 6.10. Hieruit blijkt eveneens dat er een groot verschil bestaat tussen de lengte van de grote en de kleine as van de ellips. De coördinaten van de punten uitgedrukt tegenover het nieuwe assenstelsel geven een beschrijving van de verdeling van de punten die, voor veel toepassingen, een aanvaardbare benadering kan zijn voor hun exacte relaties. Alhoewel informatie verloren gaat door de posities op de tweede as niet te gebruiken, geeft de eerste as in Fig. 6.10 een veel betere beschrijving van de variatie in de gegevens dan de tweede as die een veel kleinere spreiding in de gegevens vertegenwoordigt.
De hoeveelheid informatie die verloren gaat, hangt af van het verschil in lengte tussen de twee assen en uiteindelijk van de correlatiematrix tussen de soorten. Hoe hoger de intercorrelaties (dus ook hoe homogener de datamatrix) hoe platter de puntenwolk zal zijn. Indien geen correlaties bestaan dan zou de puntenwolk een cirkel zijn en de nieuwe assen zouden geen extra informatie samenvatten. In het geval dat de correlatie perfect zou zijn zouden de punten op een lijn liggen (vlak, hypervlak) overeenkomstig met de belangrijkste as. Het verschil in lengte tussen de eerste en de tweede as (onbestaand) is dan maximaal.

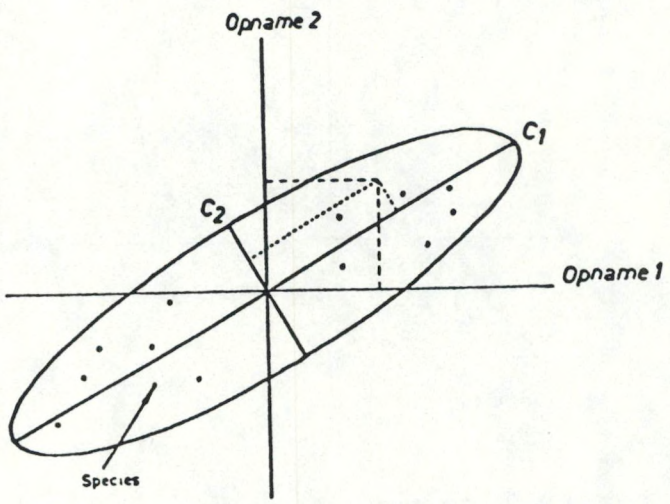


Fig. 6.10. Geometrische voorstelling van de overgang van de opnameruimte (NR=2) naar een nieuwe componentenruimte (C1, C2) door rotatie (naar Gittins, 1969).

Geometrisch bekeken komt PCA dus neer op 1) een translatie: alle punten worden getranslateerd door standardisatie, zodat het assenstelsel in het midden van de puntenwolk komt te liggen; en 2) een rotatie waarbij de onderlinge afstand tussen de punten ongewijzigd blijft maar wel zo dat de variatie langs de eerste as het grootst wordt en langs de verdere assen daalt). Bij PCA is het zo dat men steeds orthogonale assen bekomt. De assen worden geroteerd totdat de variantie van de ordinatiescores gesommeerd over alle punten en loodrecht geprojecteerd op de eerste PCA-as, maximaal is of dat de residuele som van kwadraten het kleinst is. Dit is het kwadraat van de afstand van de loodrechte projecties van de punten op de as (analoog aan de methode van de kleinste

kwadraten in regressie analyse).

Een dergelijke werkwijze resulteert dus niet in een reductie van het aantal dimensies. Maar aangezien het grootste deel van de informatie verklaard wordt door de eerste assen (gewoonlijk 2-3) spelen de overige assen slechts een geringe rol. Hierdoor draagt PCA in aanzienlijke mate bij tot de samenvatting van de originele matrix.

De ordinatie is compleet na het aangeven van de coördinaten van de punten in relatie tot de hoofdcomponenten. Gewoonlijk geeft men eveneens een aanduiding van de variatie verklaard door de nieuwe assen, dit is de lengte of het belang van de opeenvolgende componenten. Dit wordt uitgedrukt in de vorm van eigenwaarden. Deze geven per as het percentage van de totale variatie die door deze as wordt verklaard. Het is verder belangrijk om te beseffen dat, alhoewel de coördinaten van de punten verschillen tussen de oorspronkelijke assen en de geroteerde assen, de configuratie van de punten ongewijzigd blijft door transformatie. In het bijzonder blijven de hoeken en de afstanden tussen de punten onveranderd.

OPNAME ORDINATIE

Veelal bekomt men bij toepassing van PCA eveneens een ordinatie van de opnamen. Dit wordt bereikt door de gestandardiseerde scores van de soorten in een bepaalde opname te vermenigvuldigen met hun coördinaten op een bepaalde component. Sommering van deze produkten over de soorten van een opname resulteert in de coördinaat van die opname in relatie tot die component. Dit proces wordt voor alle opnamen herhaald en geeft een ordinatie van de opnamen. De coördinatenladingen van de soorten op de componenten worden hier gebruikt als gewichten (vgl. weighted averages) om de relaties tussen opnamen voor te stellen.

R- EN Q- ORDINATIES

De datamatrix tot nogtoe beschouwd, werd voorgesteld in relatie tot de opnameassen. PCA geeft dan uiteindelijk een ordinatie van de soorten (species loadings) waarvan een opname- of monsterordinatie kan afgeleid worden. Dit wordt normaal een R-analyse genoemd. Dezelfde datamatrix kan evengoed bekeken worden als de locatie van opnames in een soortsruimte. Dezelfde procedure zal nu direct een ordinatie van de monsters opleveren waarvan daarna de soortsordinatie kan bekomen worden. Dit is een Q-analyse. Het kan nu eenvoudig aangetoond worden dat, gegeven het gebruik van eenzelfde transformatie, R en Q analyses eenzelfde resultaat opleveren. De keuze tussen de analyse wordt dan vooral bepaald door de rekentijd. Zijn er minder soorten dan monsters dan neemt men beter een R analyse, dit geeft dan een lagere dimensionaliteit, en omgekeerd.

BEREKENING VIA MATRIX ALGEBRA

In de praktijk wordt een ordinatie niet grafisch uitgevoerd en niet beperkt tot een paar dimensies. In de multi-dimensionele ruimte worden de corresponderende algebraïsche bewerkingen uitgevoerd door een computer. Uit de aard van de bewerkingen kan men eveneens afleiden dat PCA in wezen een lineair model is dat algebraïsch samen te vatten is als volgt:

$$F_j = A_{1j}Z_1 + A_{2j} + \ldots\ldots + A_{NSj}Z_{NS} \quad (j = 1, \ldots, NS)$$

waarbij F_j = de j^e hoofdcomponent is van de waargenomen variabelen Z_1, Z_2, ... , Z_{NS} en A_{1j}, A_{2j}, ..., A_{NSj} de coëfficienten zijn die het relatieve belang van iedere soort of andere variabele in de afgeleide component weergegeven.

DIVERSE PCA VARIANTEN

Op de hierboven beschreven berekeningswijze van PCA bestaan heel wat varianten, waarvan de uiteindelijke resultaten uiteraard verschillen. Gezien in sommige programma's de diverse keuzes aangeboden worden en voor andere het essentiëel is om weten welk soort PCA wordt uitgevoerd, gezien de resultaten erdoor bepaald worden, proberen we hier de verschillende alternatieven kort te situeren. In ieder geval gaan we ervan uit dat op basis van de datamatrix een secundaire matrix berekend wordt die gebruikt

wordt in de eigenanalyse. Wanneer we hierin het soortsgemiddelde aftrekken dan bekomen we een "centered PCA". Dit kunnen we heel simpel als volgt voorstellen. In Fig. 6.11 zijn enkele opnamen weergegeven in een tweedimensionele soortsruimte.

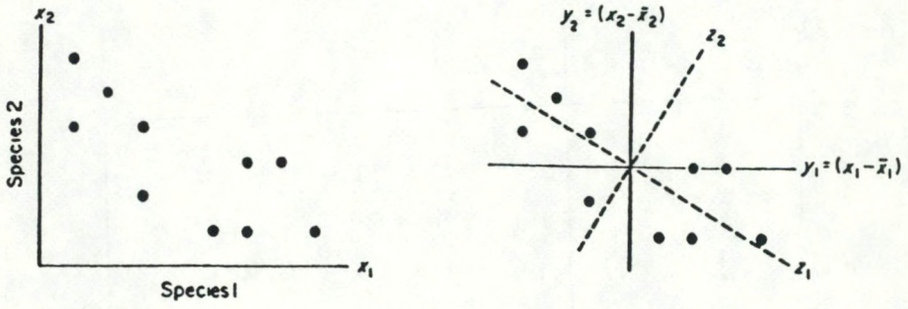


Fig. 6.11. Weergave van enkele opnames in een twee dimensionele soortsruimte en het nieuwe assenstelsel na het aftrekken van de soortsgemiddelden.

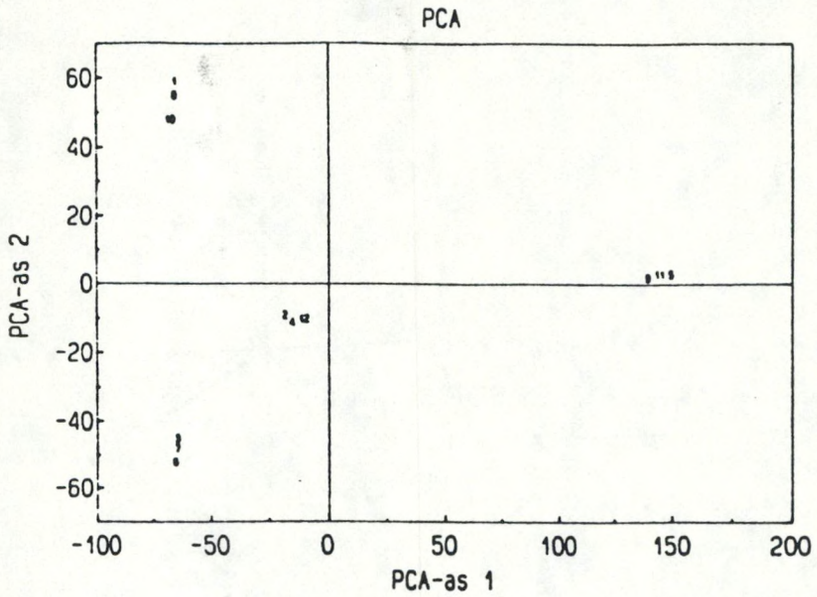
Door van elk punt het gemiddelde van de soort af te trekken bekomen we een nieuw assenstelsel met de oorsprong op de gemiddelde waarde van elke soort. Daarna volgt een rotatie rond het centroid waardoor de eerste PCA as zo gekozen wordt dat ze de grootste variatie in de punten omvat. Anders gezegd de variantie van de scores (de loodrechte projectie van de monsterpnten op de as) wordt gemaximalizeerd. De tweede PCA-as staat daar loodrecht op. In een tweedimensioneel systeem is dit triviaal maar in een meer- dimensioneel systeem niet meer. Deze as verklaart dan maximaal de overblijvende variatie. Wanneer de soorten gecentreerd en gestandardiseerd worden (zoals hierboven beschreven) dan spreken we van "centered and standardized" of simpelweg "standardized PCA". Daarnaast kunnen we PCA ook nog uitvoeren op gestandardiseerde of nietgestandardiseerde maar niet gecentreerde gegevens.

Voor discussie zie:....

PCA is in diverse paketten opgenomen. Gezien in CANOCO diverse varianten zijn opgenomen en we dan op basis van dezelfde datamatrix PCA, DCA en CCA kunnen berekenen maakt dit wel het handigste programma. In Fig. 6.12 is een voorbeeld opgenomen waarin we een gedeelte van de resulterende CANOCO listing terugvinden. In Fig. PCA10 zijn twee varianten van PCA toegepast

op eenzelfde datamatrix wat duidelijk aantoont dat de keuze van de methode een grote invloed heeft op de data.

Fig. 6.12 Resultaten van een PCA analyse met CANOCO (Figuur volgt).



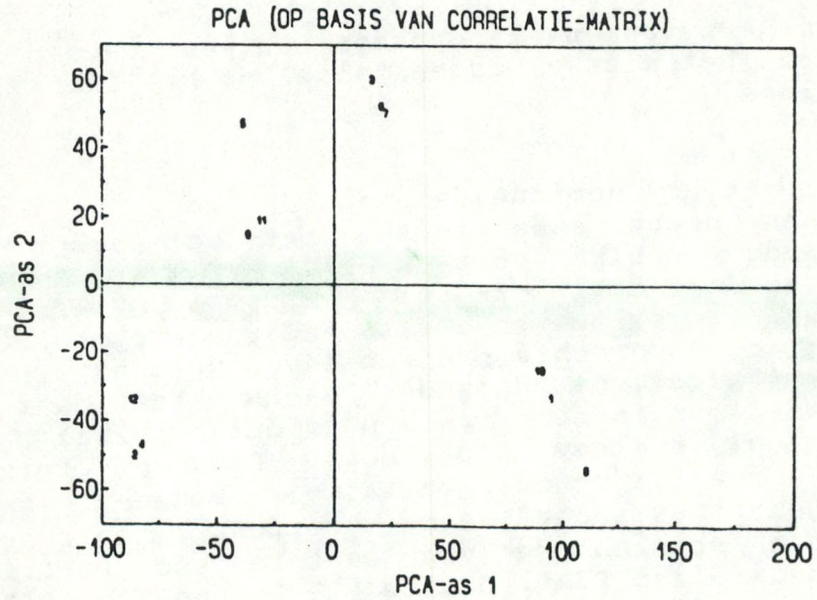


Fig. 6.13 Voorbeeld van twee verschillende versies van PCA toegepast op eenzelfde dataset.

MDSCAL (MULTIDIMENSIONAL SCALING)

MDSCAL werd ontwikkeld door Shepard (1962) en mathematisch uitgebouwd door Kruskal (1964a, b). De onderliggende filosofie van MDSCAL is simpel: gegeven een matrix met similariteiten tussen alle paren van monsters, maakt MDSCAL een configuratie van punten in een te specifiëren dimensie, zodanig dat de afstanden tussen de punten onderling (in de configuratie) maximaal overeenkomen met de rankorde van de similariteiten tussen de punten. Anders gezegd: het centraal idee van MDSCAL is het bekomen van een monotone relatie tussen de similariteiten en de afstanden in het diagram. Bovendien geeft het een maat om aan te tonen hoe goed je dit doel bereikt hebt.

Om de gedachten te vestigen kunnen we volgend simpel voorbeeld uitwerken: 5 monsters elk bestaande uit een aantal soorten. Tussen elk mogelijk paar monsters berekenen we een (dis)similariteits maat D (zie 7.1 Similariteitsindices). In totaal krijgen we voor 10 mogelijke paren een waarde (monster 1 en 2 (D_{12}), 1 en 3 (D_{13}) t.e.m. monster 4 en 5 (D_{45})). Deze dissimilariteiten kunnen we ranken van de hoogste tot de laagste:

bv. $$D_{23} < D_{34} < ..... < D_{14}$$

Onze 5 monsters (n) willen we nu in een t-dimensionale ruimte voorstellen. Laat ons deze punten X_1, ..., X_n noemen. Deze n punten in onze t-dimensionale ruimte noemen we een configuratie. Het eerste probleem is om te weten hoe goed deze configuratie de data voorstelt. Verder zullen we dan ingaan op hoe we de configuratie, die de data het best beschrijft, bekomen.

Een eerste stap om te zien hoe goed de configuratie de data voorstelt is het bepalen van de afstanden tussen de punten. De coördinaten van X_i zijn:

$$X_i = (X_{i1}, ..., X_{it})$$

en de afstand tussen twee punten is:

$$d_{ij} = \sqrt{\sum_{s=1}^{t} (X_{is} - X_{js})^2}$$

Om het verband tussen de dissimilariteiten en de afstanden te zien kunnen we een scatterdiagram maken (Fig. 6.14). In dit diagram stelt elk punt de afstand en de dissimilariteit tussen twee monsters voor. Een perfecte overeenkomst bekomen we wanneer de afstanden tussen de punten dezelfde rankorde vertonen als de dissimilariteiten of m.a.w. in ons voorbeeld:

bv. $$d_{23} < d_{34} < ..... < d_{14}$$

Grafisch betekent dit dat wanneer we alle punten in ons scatterplot verbinden, we steeds naar rechts boven gaan en nooit naar links zoals weergegeven in Fig. 6.15. Om nu de afwijking van

deze perfecte situatie na te gaan is het logisch om een curve te fitten. Dan kunnen we voor elk punt de deviatie bepalen tussen de gemeten afstand en de curve, namelijk $d_{i,j}$ (Fig. 6.15). De stress of hoe goed een gegeven configuratie de data voorsteld wordt dan gegeven door:

$$\text{stress} = S = \sqrt{\frac{\sum_{i<j} (d_{i,j} - \hat{d}_{i,j})^2}{\sum_{i<j} d_{i,j}^2}}$$

wat niet meer is dan een klassieke "residual sum of squares" (de nominator) geschaald door de som van de kwadraten van de individuele afstanden (de denominator). Een stress van 0 betekent dus een perfecte fit.

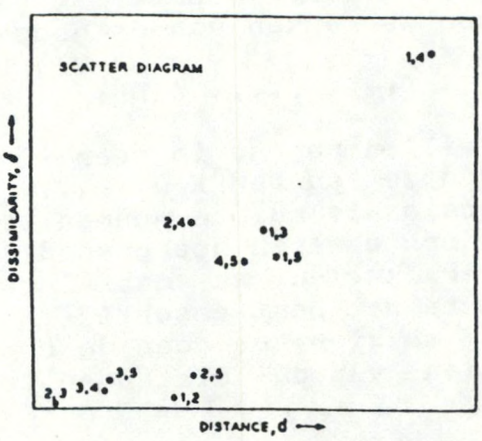


Fig. 6.14. Plot van de dissimilariteit tussen twee punten in functie van de afstand tussen deze twee punten in de configuratieruimte. (voor uitleg zie tekst).

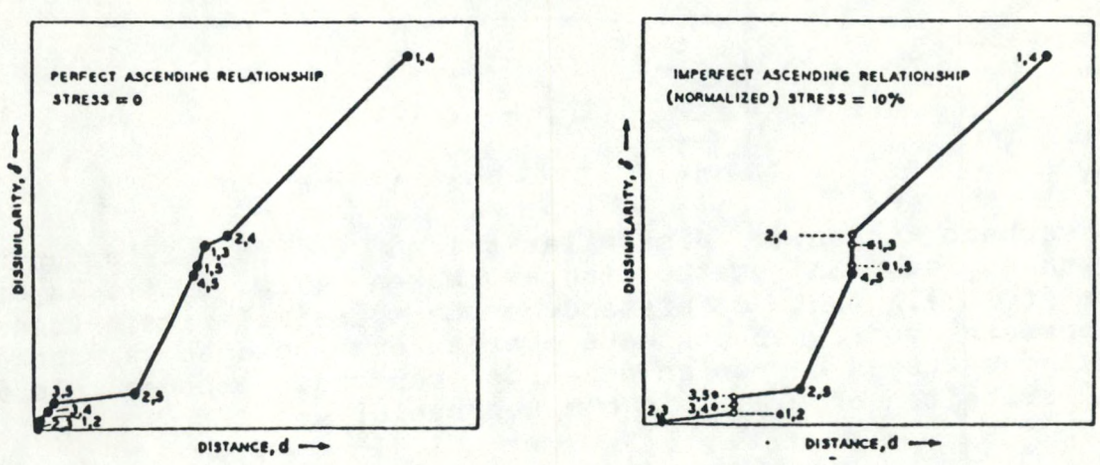


Fig. 6.15. Plot van de dissimilariteit tussen twee punten in functie van de afstand tussen deze twee punten in de configuratieruimte. a) het perfecte geval waar de ranking van beide maten identiek is; b) een "normaal geval" waar een curve gefit werd en de deviatie tussen elk punt en de curve bepaald werd.

Nu is het ook eenvoudig om de d_{ij}'s te bepalen. Het zijn deze getallen die S minimalizeren onder de constraint dat het monotoom moet zijn. .-.. De numerische methode, die gebaseerd is op een iteratief proces, wordt beschreven door Kruskal (1964b). n arbitraire punten worden geschikt (in een zogenaamde startconfiguratie). Deze configuartie kan een door het programma opgestelde random configuratie zijn, maar kan bijvoorbeeld ook bestaan uit eerder door het programma berekende ordinatiescores (coördinaten). Vervolgens wordt de euclidische afstand (of een andere maat) tussen de startconfiguratiepunten onderling berekend. Met behulp van een monotonische regressie tussen beide matrices (afstanden en eigenlijke data), waarbij de enige informatie die gebruikt wordt uit de (dis)similariteitenmatrix hun rankorde is (cf. definitie van "monotoon"), kunnen de afstanden aan de regressie gefit worden. Dit geeft de geschatte afstand tussen de configuratiepunten en de punten gedefinieerd door de similariteiten. Met een zogenaamde stressformule wordt berekend hoe groot deze afstanden zijn. De startconfiguratie wordt vervolgens iteratief aangepast (herhaling van hierboven beschreven procedure), volgens de methode van de "steepest descent" en dit tot de stress ("badness of fit") minimaal is. Dit gehele proces wordt herhaald met verschillende startconfiguraties om de invloed van de initiele keuze ervan te minimaliseren, en in verschillende dimensies om na te gaan hoeveel dimensies nodig zijn om de structuur in de data optimaal voor te stellen.

Deze techniek werd een groot succes in de psychologie ondermeer omdat ontbrekende waarden geen probleem vormen. Een bijkomend voordeel is dat deze ordinatietechniek de resultaten in niet meer dimensies dan strikt noodzakelijk weergeeft.

Een nadeel t.o.v. andere technieken is wel de relatief lange berekeningsduur en ook het feit dat soorten en monsters niet simultaan geordend worden. Het feit dat vooraf een similariteitsindex moet worden gekozen kan als een nadeel beschouwd worden, omdat opnieuw een subjectieve keuze moet gemaakt worden (die wel degelijk weerspiegeld wordt in de uiteindelijke resultaten), maar houdt anderzijds de mogelijkheid in bepaalde - aan het onderzoek eigen - accenten te leggen.

Deze techniek werd onder andere succesvol gebruikt door Field, Clark en Warwick (1982) in ecologisch onderzoek. Volgens deze auteurs is MDSCAL veel flexibeler dan de traditionele ordinatietechnieken. Verschillende varianten van deze techniek zijn geïmplementeerd in aparte programma's, maar er is ook een versie beschikbaar in het statistisch programmapakket SYSTAT.

(Fig. 6.16 volgt samen met de verschillende varianten en referenties.)

WEIGHTED AVERAGES ORDINATIE

Weighted averages is de simpelste ordinatie techniek en wordt bijgevolg ook reeds het langst gebruikt in de ecologie (Gauch, 1982). Vertrekkend van een simpele datamatrix kan een ordinatiescore S voor elke soort berekend worden op basis van de volgende formule:

$$S_i = \frac{\sum X_{ij} W_j}{\sum X_{ij}}$$

met X_{ij} het aantal van soort i in staal j en W_j is een gewicht die we aan het monster toekennen. Dit gewicht kan bijvoorbeeld een bepaalde gemeten abiotische factor zijn. De ordinatiescore van de soort kunnen we dan ook interpreteren als een soort indicatorwaarde. In Tabel 6.1 is een simpele dataset weergegeven waar op 5 monsterlocaties 5 soorten werden aangetroffen. Op de 5 plaatsen werd een omgevingsfactor gemeten.

	monster					$\frac{\sum x_{ij} W_j}{\sum x_{ij}}$
	1	2	3	4	5	
soort						soortscore
1	5	4	3	2	1	2.33
2	0	1	0	3	10	4.57
3	3	7	2	8	1	2.48
4	10	2	0	0	1	1.46
5	0	0	3	2	8	4.38
abiotische factor						
	1	2	3	4	5	

Tabel 6.1. Overzicht van een hypothetische dataset. Op 5 monsterplaatsen werden 5 soorten aangetroffen en werd de waarde van een abiotische factor gemeten. De soortscores berekend volgens weighted averages (met de waarde van de abiotische factor als gewicht) zijn weergegeven.

Op basis van de resulterende soortscores kunnen we de soorten rangschikken langs de gemeten gradient (S4 < S1 < S3 < S5 < S2)? Pas op voor de verdeling van de stalen over de gradient (cfr. Ter Braak, 1987 p. 62)

Op basis van soortsgewichten kunnen ook monsterscores worden berekend.

RECIPROCAL AVERAGING (RA) OF CORRESPONDANCE ANALYSIS (CA)

RA werd door verschillende wetenschappers tegelijkertijd ontworpen zonder dat ze weet hadden van elkaar. Het is een extensie van "weighted averages" die door Whittaker (1967) in

direct gradient analysis werd gebruikt. Hill (1979) werkte uiteindelijk het algoritme uit dat geïntroduceerd werd in de ecologie en nu algemeen gebruikt wordt. Het is gebaseerd op de volgende redenering. Wanneer we op meerdere punten een omgevingsvariabele en de soortensamenstelling gemeten hebben dan kunnen we voor elke soort zijn indicator waarde bepalen door het gewogen gemiddelde te maken van de waarde van de omgevingsvariabele in de monsterpunten waar de soort voorkomt. Op basis daarvan kunnen we de soorten rangschikken. Als de indicator waarden van de soorten gekend zijn kunnen we op analoge manier de monsters ordenen. Toegepast op de matrix uit Tabel 6.1 geeft dit de volgende matrix:

	monster					
	1	2	3	4	5	
soort						soortscore
4	10	2	-	-	1	1.46
1	5	4	3	2	1	2.33
3	3	7	2	8	1	2.48
5	-	-	3	2	8	4.38
2	-	1	-	3	10	4.57
abiotische factor						
	1	2	3	4	5	

Tabel 6.2. Data matrix uit Tabel 6.1 waarin de soorten geordend zijn volgens de soortscores.

Wanneer de soorten reageren op de omgevingsfactoren volgens een Gauscurve dan vinden we een diagoname structuur terug in de data. Deze methode is slechts toepasbaar wanneer we op voorhand weten welke omgevingsvariabelen van belang zijn. Met CA willen we nu proberen de onbekende onderliggende omgevingsgradient te vinden door niet één keer maar meerdere keren de gewogen gemiddeldes te bepalen en dit dan bovendien niet voor de soorten of de monsters afzonderlijk maar voor beide samen.

Deze ordinatietechniek vertrekt hiervoor van arbitraire monsterscores. Die monsterscores worden dan gebruikt om "gewogen soortscores" te berekenen. De monsterscore wordt vermenigvuldigd met het aantal van een soort in dit monster. Dit gebeurt voor elk monster. De som van die produkten gedeeld door het aantal monsters waarin deze soort voorkomt is dan de gewogen soortscore. Die worden uiteraard voor alle soorten berekend. Die worden dan op hun beurt gebruikt om nieuwe monsterscores te berekenen en vice versa (vandaar de term "reciprocal averaging"). Er komt echter nog één extra probleem. Door steeds maar gemiddelden te nemen wordt de range van de scores steeds kleiner. Om dit te voorkomen worden na iedere berekening of de soorts- of de monsterscores uitgedrukt op een schaal van 0 tot 100. Dit gehele proces wordt iteratief herhaald tot de scores stabiliseren en niet langer beïnvloed zijn door de initiële keuze van arbitraire soortscores. Er worden dus tegelijkertijd monsters en soorten geordend. In tabel 6.3 werken we een hypothetisch voorbeeld uit, waarbij voor de eenvoud 0 of 1 wordt gebruikt als dichtheid.

monster	dichtheden								monsterscores			
									(a)	(b)	(c)	
1	1	0	0	1	1	0	0	1	100	52.5	55	44.3
2	0	1	1	0	0	1	0	1	0	37.5	0	36.2
3	1	1	0	0	0	1	1	0	100	65.0	100	63.4
4	1	1	1	1	1	0	0	1	0	43.3	21	39.3
5	1	1	0	1	0	0	0	1	100	56.7	70	47.2
6	1	0	0	0	1	0	0	0	0	46.7	33	46.0
(1)	60.0	50.0	0.0	66.7	33.3	50.0	100.0	50.0				
(2)	55.8	47.8	10.5	48.7	36.3	50.0	100.0	36.5				
.												
.												
.												
(11)	31.8	56.5	48.4	19.7	10.0	86.0	100.0	32.7				
(11a)	24.0	52.0	42.0	11.0	0.0	84.0	100.0	25.0				

(a) = arbitraire monsterscores
(1) = gewogen soortscores$_j$: gemiddelde$_i$ (dichtheid$_{i,j}$ * monsterscore$_i$) met i het rijnummer en j het kolomnummer.
(b) = gewogen monsterscores : gemiddelde$_j$ (dichtheid$_{i,j}$ * soortscore$_j$).
(c) = gestandaardiseerde monsterscores : 100 * (monsterscore$_i$ - Min$_i$) / spreidingsbreedte$_i$.

Tabel 6.3: Iteratieve berekening van RA scores op eerste as m.b.v. een hypothetisch voorbeeld (dichtheden 0 en 1).

De resulterende scores geven een maximale spreiding en ze maken de eerste RA-as uit. De tweede as wordt bekomen via eenzelfde iteratieproces maar met één extra stap namelijk dat de scores voor de tweede as ongecorreleerd gemaakt worden met die van de eerste as. De eventuele korrelatie met lagere assen wordt nu echter als bijkomende stap voor de berekening van de monsterscores telkens opnieuw teniet gedaan door een gewogen lineaire regressie tussen de scores van de vorige as en de nieuwe scores te berekenen. De resulterende scores zijn de residuals van deze regressie ($U_j' = U_j - U_j^*$, met U_j^* de via regressie geschatte waarde van U_j; voor details zie ook Gauch, 1982).

DETRENDED CORRESPONDENCE ANALYSIS

PROBLEMEN MET RA

De twee voornaamste fouten van RA (ook bij PCA), nl. het boogeffect (ook wel hoefijzereffect genoemd) en de contractie van de schaal aan de uiteinden van de assen worden mathematisch weggewerkt.

Het boogeffect is het gevolg van het feit dat vooral de tweede as (soms ook hogere assen) een duidelijk verband vertoont met de eerste as, alhoewel een (lineaire) correlatie afwezig is (de positieve correlatie aan één zijde van de boog en de negatieve correlatie aan de andere zijde van de boog geven samen een netto-correlatie van nul). Uiteindelijke oorzaak voor dit belangrijke artefact is de structuur van de data. In de ecologie is het veelal zo dat soorten langs ecologische gradiënten min of meer klokvormige

(Gauss-curven) respons curven vertonen. Fig. 6.17 geeft hiervan een uitstekend voorbeeld. Hieruit blijkt eveneens dat afhankelijk van de lengte van het segment van de ecologische gradiënt die bemonsterd wordt een lineaire of niet-lineaire data structuur te voorschijn kan komen.

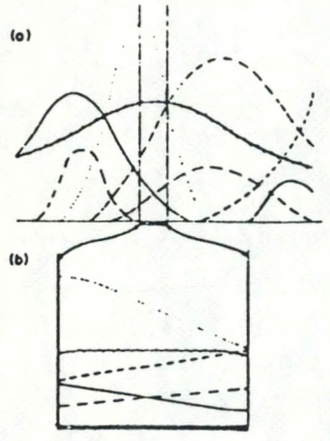
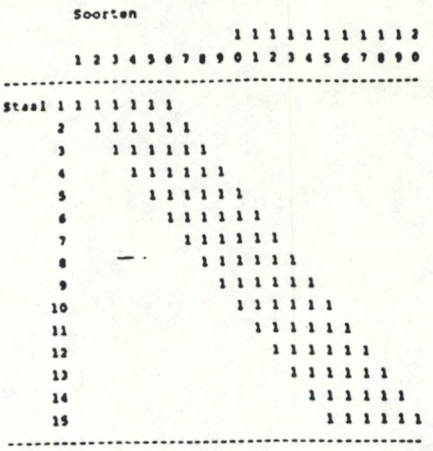
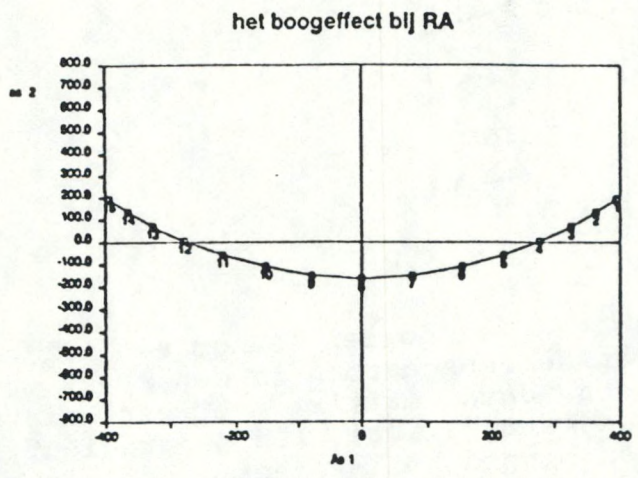


Fig. 6.17. (a) Het gedrag van 8 soorten langs een milieugradiënt. Bemonstering van de levensgemeenschappen in een reeks punten langs de hele gradiënt geeft een data matrix met een niet- lineaire structuur. (b) Vergroot segment uit (a). Als de gradiënt bemonsterd wordt op een aantal punten binnen het segment, dat zo kort is dat de respons curven van de soorten (bijna) lineair zijn, dan is de resulterende data matrix (benaderend) lineair van structuur (naar Pielou 1984).

In de meeste gevallen echter verkrijgen we niet-lineaire respons curven: soorten verschijnen, bereiken een piekwaarde en verdwijnen weer. Het effect hiervan op RA en PCA op een fictief voorbeeld wordt geïllustreerd in Fig. 6.18. De 15 stalen met presentie-absentie gegevens hebben een regelmatige data-structuur. Het resultaat van een ordinatie van stalen genomen langs een lineaire gradiënt zou een lineair patroon moeten vertonen. Zowel PCA als RA vertonen hier echter duidelijk het boogeffect, PCA meer uitgesproken dan RA. RA heeft ten opzichte van PCA een beter resultaat: de boog is minder uitgesproken, maar de ligging op de 2de as is zonder betekenis.
Ook is duidelijk te zien in Fig. 6.18 hoe de stalen op de uiteinden van as 1 dichter bij elkaar liggen dan in het midden. Deze variatie in onderlinge afstand correspondeert niet met verschillen in de data: de opeenvolgende stalen verschillen even veel van elkaar. Er is dus ruimte voor verbetering.





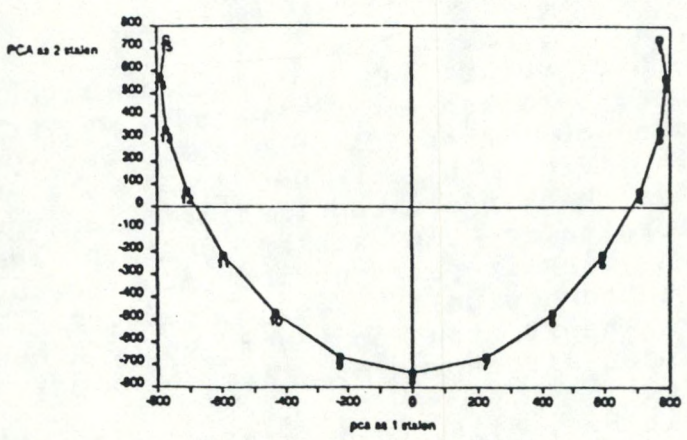


Fig. 6.18. (a) Voorbeeld van een fictieve data matrix met presentie- absentie gegevens waarin de opeenvolgende stalen even veel van elkaar verschillen. Er is dus slechts één natuurlijke gradiënt.
(b) PCA- ordinatie van de 15 stalen en (c) RA-ordinatie van dezelfde stalen. Beide ordinaties vertonen het boogeffect. Vooral bij de RA-ordinatie liggen de stalen aan de uiteinden van de as 1 dichter bij elkaar dan in het midden. Volgens de originele regelmatige data structuur is hier geen enkele reden voor.

MOGELIJKE OPLOSSINGEN

DCA is een verdere ontwikkeling en verbetering van Correspondentie Analyse (CA) (beter bekend als Reciprocal Averaging (RA)) en werd ontworpen door Hill (1979a, Hill & Gauch 1980). Het vlakt de misleidende boog af en het corrigeert voor de contractie in schaal aan de uiteinden van de RA-ordinatie. Om het boogeffect te vermijden moet er voor gezorgd worden dat de op de eerste as volgende assen geen enkele systematische relatie vertonen met de eerste as. Eén mogelijkheid om dit te doen is er voor zorgen dat voor ieder punt langs de eerste as de gemiddelde waarde van de opeenvolgende assen nul benadert (Fig. 6.19). Daarvoor wordt de eerste as in een aantal (zelf te kiezen) segmenten verdeeld en scores van de tweede as worden aangepast zodat ze een gemiddelde score van nul hebben. Dit 'detrending' wordt toegepast bij iedere iteratie, uitgezonderd bij convergentie, dan worden de scores van de stalen berekend als gewogen gemiddelden van de scores van de soorten (zoals in RA). Mathematische details vindt men terug in Hill (1979a).

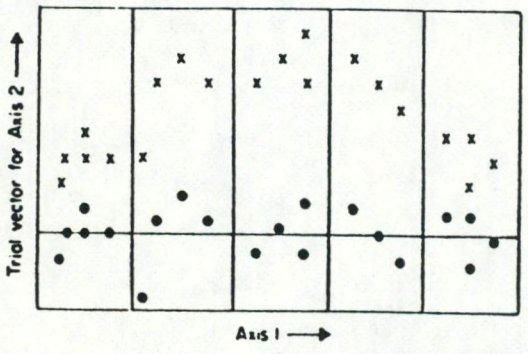
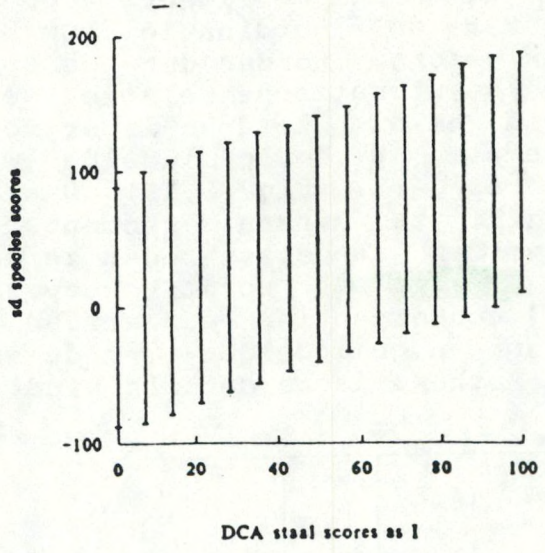
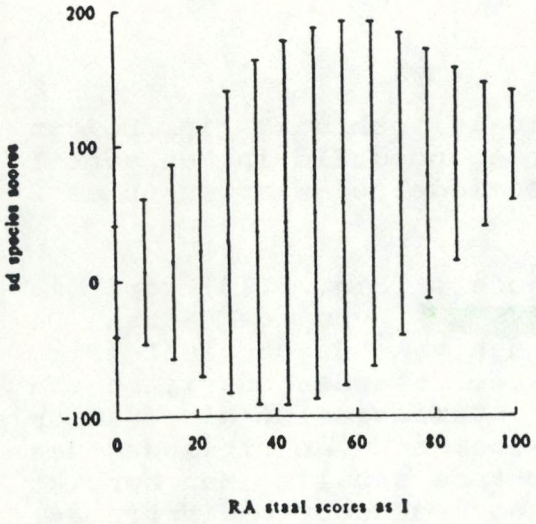


Fig. 6.19. De methode van detrending zoals gebruikt in DCA in DECORANA. De gradiënt langs as 1 wordt opgedeeld in een aantal segmenten. Binnen elk segment worden de gemiddelde waarden op as 2 gelijk (naar Hill & Gauch 1980).

De andere tekortkoming van RA (zie hoger) (Fig. 6.18) wordt in DCA weggewerkt door herschaling ('rescaling') van de assen. De contractie van de schaal aan de uiteinden van de gradiënt (Fig. 6.20) wordt weerspiegeld in een gereduceerde standaarddeviatie van de soortscores in het staal. M.a.w. de standaardafwijking van de scores van de soorten in een staal is kleiner aan de uiteinden van de gradiënt dan in het midden. Een betere schaling kan bereikt worden door die delen van de gradiënt (langs de soortordinatie-as) waar de standaardafwijking klein is uit te spreiden en die delen waar deze groot is samen te trekken. Voor een meer mathematische uitwerking verwijzen we opnieuw naar Hill (1979a). Deze aanpassing van de scores heeft het gewenste effect, omdat de eis van uniformiteit van de intra-staal spreiding van de species scores gelijkwaardig is met vragen naar uniformiteit van de intra-species spreiding van de staal scores (zoals uit Fig. 6.20 blijkt, waar voor elke positie langs de gradiënt de lengte van de vertikale lijnen voor de soorten ongeveer gelijk is aan de lengte van de horizontale lijnen van de stalen). Na herschaling is de gemiddelde spreiding binnen elk staal van de species scores op alle punten van de ordinatie-as van de stalen ongeveer gelijk (Fig. 6.21). Wordt deze standardisatie gelijk gesteld aan 1 dan bekomt men een

standardisatie, waardoor het gemiddelde abundantie-dominantie profiel van de soorten een gemiddelde standaardafwijking (SD) heeft





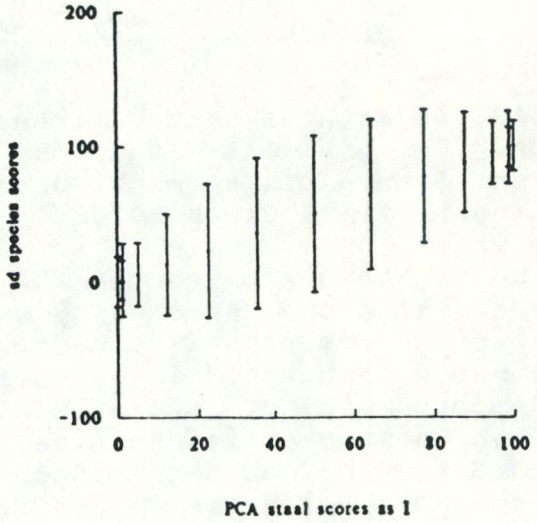
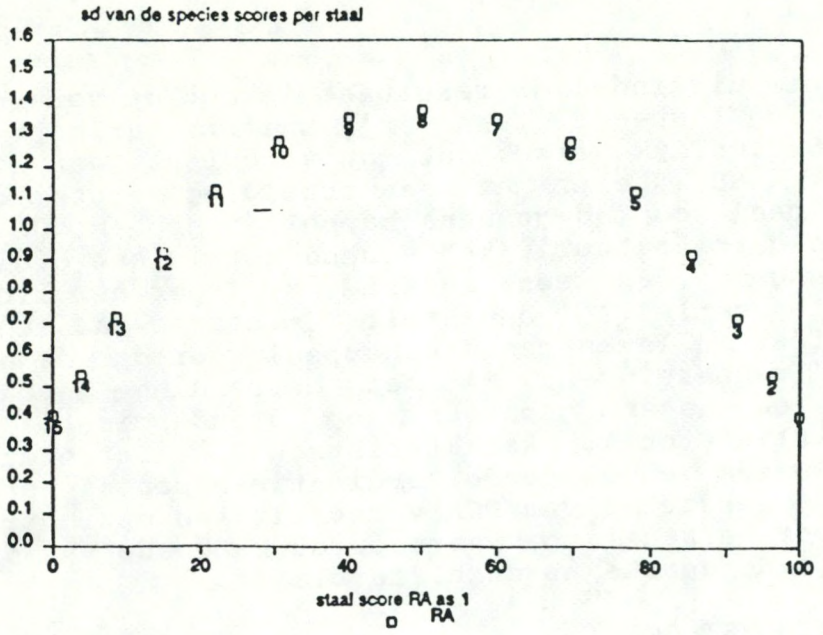


Fig. 6.20. De verandering van de standaardafwijking van de ordinatie-scores van de soorten per staal in relatie tot de ordinatie-scores van de stalen. De data matrix is deze uit fig. 2. (a) DCA; (b) RA; (c)PCA.
De standaardafwijking van de soortscores binnen ieder staal is kleiner aan de uiteinden dan in het midden bij PCA maar vooral bij RA. De bedoeling van de herschaling binnen DCA is elke systematische relatie tussen deze standaardafwijkingen en de ordinatiepositie te verwijderen (a).



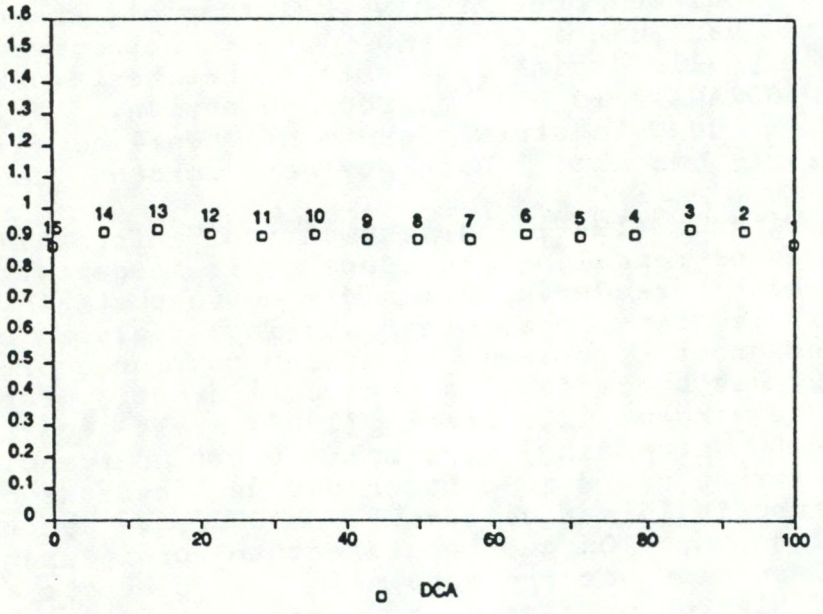


Fig. 6.21. De spreiding van de soortscores binnen elk staal in relatie tot de positie langs de eerste ordinatie-as (data van Fig. 6.18). Hieruit blijkt dat met DCA de spreiding van de soortscores nagenoeg constant is, terwijl deze bij RA in het midden van de as veel groter is dan aan de uiteinden.

gelijk aan 1. Het uiteindelijk resultaat is dat de verandering in soortensamenstelling ('species turnover') uniform verloopt langs de ordinatie-as. Gewoonlijk verschijnt en verdwijnt een soort over een range van ca. 4 SD. Stalen met een scheiding groter dan 4 SD zullen algemeen geen soorten gemeenschappelijk hebben. De schaling in DCA is bijgevolg in "natuurlijke" eenheden. Als een ordinatie-as bvb. 5.4 SD lang is en een andere as (voor dezelfde of een verschillende data set) 4.1, dan stelt de eerste as een langere coenocline (gradiënt in levensgemeenschappen) voor dan de andere.

De besproken aanpassingen, RA met detrending gevolgd door herschaling van de assen gebaseerd op standardisatie van de intra-staal variantie tot 1, karakteriseren DCA. Detrending wordt toegepast op de tweede en volgende ordinatie-assen en herschaling op alle assen. De eerste as van DCA verschilt ten opzichte van die van RA slechts door herschaling, waardoor de rangorde van de stalen (of soorten) langs de eerste as dezelfde blijft.

TESTEN EN PROBLEMEN MET DCA

Uitgebreide testen met RA, DCA en nonmetrische multidimensional scaling (MDSCAL) (Gauch et al. 1981) hebben aangetoond dat DCA vrijwel nooit resultaten geeft die moeilijker interpreteerbaar zijn dan RA en MDSCAL. Nochtans heeft ook DCA tekortkomingen (Hill & Gauch 1980:56). De problemen doen zich vooral voor bij aanwezigheid in de data-matrix van uitbijters en grote heterogeniteiten. Uitbijters worden bij alle ordinatietechnieken het best verwijderd. Het programma DECORANA is hierop voorzien. Bij grote discontinuïteiten in de datamatrix worden de onderlinge afstanden niet steeds correct geschat. Datamatrices worden dan beter opgedeeld.
Pielou (1984) en Minchin (1987) waarschuwen voor informatieverlies bij de detrending en of rescaling procedure zoals toegepast in DCA. Ter Braak (1987) raadt veranderingen aan die de robuustheid van DCA kunnen verhogen. 1/ de niet-lineaire herschaling zoals in DECORANA uitgevoerd is geen groot probleem, zodat een routine gebruik niet aan te raden is; 2/ het boogeffect is wel een ernstig probleem en dient verwijderd te worden. Ter Braak (1987) stelt een nieuw en minder "aggressief" detrending algoritme voor: detrending door polynomialen. De eerste as is zo gebogen dat de tweede as ongeveer een kwadratische functie is van de eerste, de derde as een kubische functie van de eerste enz.. Om het boogeffect te verwijderen mag de tweede as niet alleen niet gecorreleerd zijn met de eerste as (xi), maar moet ze ook niet gecorreleerd zijn met het kwadraat van de eerste (xi2) en bij voorkeur ook niet met de derde macht (xi3). In tegenstelling tot detrending d.m.v. segmenten verwijdert detrending d.m.v. polynomialen alleen de specifieke fouten van RA die nu theoretisch begrepen zijn. Deze nieuw methode is ingebouwd in CANOCO (ter Braak 1987b).
Wartenberg, Ferson & Rohlf (1987) zien de "boog" als een kenmerk eigen aan de data-structuur en dus niet als een artefact. Wartenberg et al. (1987) stelt ook de zin van de herschaling in vraag en besluit uit een vergelijkende studie van enkele ordinatietechnieken dat DCA geen verbetering vormt op RA en zeker niet boven alle technieken steeds te verkiezen is. Desondanks blijft DCA in de ecologie een veel gebruikt techniek. Het efficiënte algorithme, in termen van uitvoeringssnelheid en benodigde geheugenruimte, is hier niet vreemd aan.

Verder blijft het probleem bij alle ordinatietechnieken (Hill & Gauch 1980:57):
'The interpretation of results remains a matter of ecological insight and is improved bij field experience and bij integration of supplementary environmental data for the vegetation sample sites'.
Vooral aan dit laatste aspect wordt in een verdere ontwikkeling van DECORANA, nl. CANOCO wel extra aandacht besteed.

ENKELE BELANGRIJKE PROGRAMMA-SPECIFICATIES VAN DECORANA

De standaardanalyse (default waarden) in DECORANA geniet zeker de voorkeur wanneer een data matrix voor de eerste keer wordt geanalyseerd. Later kunnen variaties aangebracht worden. Een belangrijk hulpmiddel voor interpretatie blijft het uitschrijven van de stalen en soorten in volgorde van de ordening op de eerste as (en vervolgens tweede, derde en eventueel volgende as). Het directe contact met de data matrix blijft op die manier behouden. Bij ORDIFLEX is dit ingebouwd, bij DECORANA niet.

Uitbijters

Na een eerste analyse van de gegevens (bvb. met DECORANA) kan het nodig zijn om uitbijters weg te laten. DECORANA voorziet na inlezen van de data set, na de input van alle gegevens, de mogelijkheid om stalen weg te laten.

Transformaties

Vooral bij gebruik van metingen (bvb. dichtheid in aantallen, drooggewichten) met grote verschillen tussen minima en maxima kunnen dominante soorten het ordinatie-resultaat heel sterk beïnvloeden. Om dit te voorkomen worden veelal transformaties toegepast (zie hoger). DECORANA (o.a. ook ORDIFLEX,CANOCO) bieden de mogelijkheid om gegevens vooraf te transformeren. Bij DECORANA worden de waarden en hun nieuwe waarden per twee ingegeven (maximum = 46 koppels in stijgende volgorde).
Een voorbeeld van input in DECORANA:

```
0.1     1.0    (= waarde   + waarde na transformatie)
2       2
5       3
10      4
20      5
-1      0      (beëindigt de transformatie)
```

Voor andere tussenliggende waarden gebeurt een lineaire interpolatie. Dus 6.9 wordt omgezet tot 3.0+ (6.9-5.0)*(4.0-3.0)/(10.0-5.0)=3.38. Waarden groter dan de opgegeven maximum waarde (hier 20) worden omgezet naar de grootste waarde (dus naar 5).
Waarden die getransformeerd moeten worden dienen in stijgende volgorde ingegeven te worden. Als dit niet zo is, wordt de vraag tot ingave van de transformatie opnieuw gesteld. Dit kan bijgevolg dus ook gebruikt worden om tikfouten te corrigeren.

"Downweighting" van zeldzame soorten

In bepaalde toepassingen kunnen zeldzame soorten de analyse verwringen. Ook leveren zeldzame soorten veelal weinig extra informatie in ordinaties, waar in de eerste plaats de hoofdlijnen

van variatie in de data matrix weergegeven worden. DECORANA voorziet in de mogelijkheid om zeldzame soorten naar (beneden) te wegen. Als AMAX de frequentie is van de algemeenste soort dan is het effect van de weging een reductie van de bedekking (in relatie tot de frequentie) van de soorten die minder frequent zijn dan AMAX/5. Soorten algemener dan AMAX/5 worden niet naar beneden gewogen.

Laat:

b_{ij}: de bedekking van soort j in opname i
x_{ij}: nieuw score als fj kleiner is dan of gelijk aan AMAX/5

(1) Presentie-absentie gegevens (0,1):

fj : frequentie van soort j
xij= bij*fj/(AMAX/5) indien fj > AMAX/5 is xij=bij, bij is hier =1

(2) Kwantitatieve gegevens

fj= (Σ bij)2/ Σ b2ij : een kwantitatieve maat voor het aantal keer dat een soort voorkomt
xij= bij*fj/(AMAX/5) indien fj > AMAX/5 is xij=bij

Herschaling van de assen en het aantal segmenten.

Zoals eerder opgemerkt worden de assen herschaald zodat de variantie van de soortsscores in een staal ongeveer gelijk worden aan 1. Het gewenste resultaat wordt gewoonlijk niet in één keer bereikt; in het default geval wordt de herschaling 4 keer toegepast. Hill (1979a) raadt aan om deze default waarde niet te veranderen.
De segmenten worden gebruikt bij detrending van de tweede en hogere assen. De default waarde (26) heeft bewezen bevredigende resultaten te geven. Hill (1979a) raadt aan deze waarde niet te veranderen. Men kan ook een drempel voor de herschaling op geven ("rescaling threshold"). Als de drempelwaarde op t gezet wordt, dan zullen assen die een lengte hebben van minder dan t SD niet herschaald worden. De default waarde is t=0. De ervaring leert dat de bovenvermelde default waarden meestal volstaan.

HET ORDINATIE RESULTAAT

Alle ordinatie-scores zijn vermenigvuldigd met 100. Dus een score van 534 moet geïnterpreteerd worden als 5.34 SD. De laagste staal score op iedere as is gelijk gesteld aan 0. Staalscores zijn gelijk aan de gemiddelde scores van de soorten voorkomend in dat staal; de scores van de soorten hebben een grotere spreiding dan de scores van de stalen. De laagste score van de soorten is steeds kleiner dan 0.
Slechts de eerste 4 assen worden uitgerekend. Hogere assen bevatten immers gewoonlijk weinig extra informatie en vergen alleen maar (kostbare) rekentijd. De eigenwaarde, een maat voor de variantie langs de betreffende as, wordt per as vermeld, maar mag niet absoluut geïnterpreteerd worden zoals in PCA (waar alle assen uitgerekend worden). Algemeen zijn assen die een eigenwaarde hebben die veel kleiner is dan de topwaarde (de eerste as) van weinig betekenis. Ruwweg is de eigenwaarde van een as proportioneel tot

het kwadraat van de lengte van de staal ordinatie.
De opgegeven lengte van de gradiënt is de lengte van de ordinatie van de stalen. Zelfs zonder variatie in de data (beta- diversiteit = 0), kan de lengte van de ordinatie van de soorten 3 à 4 SD bedragen. Een ordinatie van de stalen van 4 SD daarentegen betekent dat soorten voorkomend aan één kant van de gradiënt (bijna) volledig afwezig zijn aan het andere uiteinde en vice versa.

CANONICAL CORRESPONDENCE ANALYSIS (CANOCO)

Alle vorige ordinatiemethoden vormen de basis voor een "indirect gradient analysis". Toch is een "direct gradient analysis" een betere methode om de relaties tussen het voorkomen van een soort en omgevingsvariabelen na te gaan. Dit kan ofwel door het grafisch weergeven van de abundantie van een soort in functie van een omgevingsvariabele of door het berekenen van correlaties of door simpele of multipele regressie-analyse. De grote hoeveelheid werk bij redelijk wat soorten of de onderliggende assumpties van regressie-analyse zijn de oorzaak dat deze manier van werken weinig gebruikt wordt. Een simpele methode is daarom nodig om de relaties tussen veel soorten en omgevingsvariabelen te analyzeren en te visualizeren. Canonical correspondance analysis (CCA) is precies ontwikkeld om aan deze behoefte te voldoen. Het is een eigenvector ordinatie methode die de variatie in een gemeenschap visualiseert zoals in een klassieke ordinatie, maar daarnaast een "direct gradient analysis" uitvoert om de gemeenschapssamenstelling rechtstreeks in verband te kunnen brengen met de gekende variatie in het milieu en dit eveneens visueel voor te stellen. Deze multivariate techniek, niet te verwarren met Canonical Correlation Analyse, werd door Ter Braak (1985) ontwikkeld en is een rechtstreekse verbetering van correspondentie-analyse (RA), maar de ordinatieassen worden gekozen in functie van de milieuvariabelen (waarvan de term "correspondentie" afkomstig is: de overeenkomsten tussen milieu- en soortvariabelen worden benadrukt). De ordinatieassen moeten namelijk lineaire combinaties zijn van de milieuvariabelen (die zowel kwantitatief als nominaal mogen zijn), hetgeen dus een meer rechtstreekse benadering is dan bijvoorbeeld bij RA, waar we de ordinatieassen achteraf korreleren met de omgevingsfactoren. Er kunnen net zoveel assen berekend worden als er milieufactoren zijn.

De afleiding van deze methode als een benadering van de statistisch veel rigoureusere Gaussiaanse canonische ordinatie wordt beschreven door Ter Braak (1986b). Een veel simpelere afleiding gebaseerd op "weighted averaging" wordt hier besproken en is gebaseerd op Ter Braak (1987a,b). Ook in de handleiding bij het programma (Ter Braak, 1986a) zijn veel details evenals de mathematische achtergrond te vinden.

Om de gedachten te vestigen vertrekken we van een hypothetisch voorbeeld zoals weergegeven in Fig. 6.22. Hierin zijn de responscurven van een viertal soorten weergegeven langsheen een omgevingsgradient, bijvoorbeeld vochtigheid. D verkiest vochtiger situaties dan A.

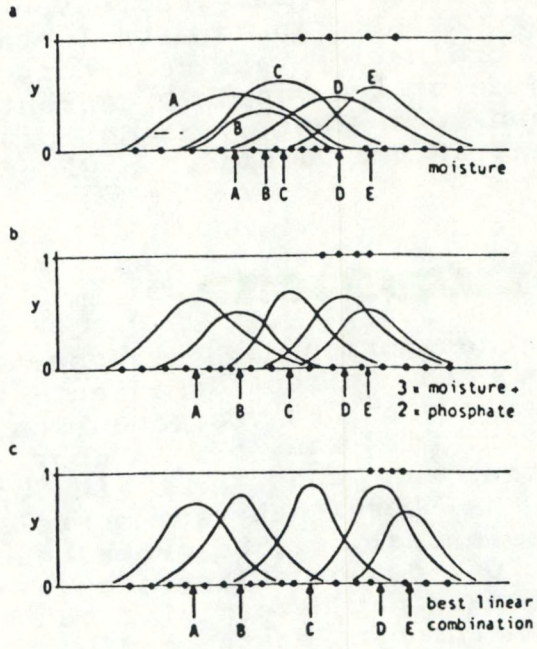


Fig. 6.22. Hypothetisch voorbeeld van unimodale respons curves van 4 soorten (A-D) langs gestandardizeerde omgevingsvariabelen die een verschillende mate van scheiding van de soortscurven geven. (a) vochtigheid; (b) een lineaire combinatie van vocht en fosfaat; (c) de beste lineaire combinatie van omgevingsvariabelen zoals gekozen door CCA. De monsters zijn weergegeven met stippen. Voor soort D is het voorkomen weergegeven: bij y=1 is de soort aanwezig, bij y=0 is de soort afwezig. (naar ter Braak, 1987).

De vraag is nu: hoe goed verklaart vochtigheid de soortsgegevens. Dit kunnen we als volgt bepalen via de methode van "weighted averaging". Voor elke soort kan een score berekend worden door het gewogen gemiddelde te nemen van de vochtgetallen van de diverse monsterpunten. Dit kan als volgt:

$$\mu_k = \sum_{i=1}^{n} y_{ik} * x_i / y_{+k}$$

met μ_k het gewogen gemiddelde van soort k, x_i het vochtgetal van monster i, y_{ik} de abundantie van soort k in monster i and y_{+k} de som van de dichtheden van soort k. In het geval van binaire data is het gewogen gemiddelde niet meer dan het gemiddelde van de vochtgetallen waar de soort aanwezig is. Het gewogen gemiddelde geeft een eerste indicatie van waar de soort voorkomt langs de vochtigheidsgradiënt zoals weergegeven met pijlen in Fig. 6.22. Een maat van hoe goed vochtigheid de soortsgegevens verklaart is de spreiding van de gewogen gemiddelden. Wanneer de spreiding groot is

dan scheidt vochtigheid de soortscurven behoorlijk en verklaart vochtigheid goed het voorkomen van de soorten. Is die spreiding laag dan hebben we juist het tegengestelde. Gezien de omgevingsvariablen gestandardizeerd zijn naar gemiddelde = 0 en variantie = 1 kunnen we door het berekenen van de spreiding van de gewogen gemiddelden langs al de beschouwde omgevingsvariabelen de "beste" variabelen uitselecteren. Stel nu dat vochtigheid de beste variabele is. Toch kunnen we een betere variabele bedenken die een combinatie is van twee andere. In ons hypothetisch voorbeeld kunnen we (willekeurig) de combinatie 3 * vochtgehalte plus 2 * fosfaat gehalte nemen. Uit Fig. 6.22 kunnen we afleiden dat de spreiding van de gewogen gemiddelden van de vier soorten een grotere spreiding vertoont dan bij vochtigheid alleen. De responscurves zijn smaller en de waarnemingen van soort D liggen dichter bij elkaar. Het is dus met andere woorden essentieel om niet alle omgevingsfactoren afzonderlijk, maar precies de lineaire combinaties ervan te bekijken onder de vorm van:

$$Xi = B1*Zi1 + B2*Zi2+ \ldots + BpZip$$

waarbij Zij de waarde van de jde omgevingsvariabele is op monsterpunt i, Bi het gewicht dat aan die variabele toegekend wordt en Xi is de waarde van de samengestelde omgevingsvariabele die aan punt i toegekend wordt. CCA is nu de techniek die de beste lineaire combinatie van omgevingsvariabelen selecteert die de dispersie van de soortscores maximaliseert of m.a.w. CCA kiest de beste gewichten voor de omgevingsvariabelen. Door deze restrictie, namelijk de sitescores zijn lineaire combinaties van de gemeten omgevingsvariabelen, in te bouwen in het "two-way weighted averaging" algoritme van RA krijgen we het algoritme van CCA. De werkwijze is dus als volgt:

1) kies willekeurige maar ongelijke initiële sitescores

2) bereken de speciesscores op basis van de gewogen gemiddelden van de sitescores

3) bereken nieuwe sitescores als gewogen gemiddelden van de soort-scores

4) bereken de regressie-coefficiënten door een gewogen multipele regressie van de sitescores met de omgevingsvariabelen. (de gewichten zijn de totalen per monsters)

5) bereken nieuwe sitescores als de gefitte waarden van de regressie berekend in de vorige stap.

6) centreer en standardiseer de sitescores

7) stop bij convergentie (wanneer de nieuwe sitescores voldoende dicht bij de vorige liggen of ga terug naar stap 2

Met uitzondering van stap 4 en 5 is de werkwijze dus volledig analoog aan RA.

De overige assen kunnen eveneens op dezelfde manier als bij RA berekend worden.

Het uiteindelijke resultaat kan op diverse niveaus geanalyseerd worden. Vooreerst proberen we de ordinatieassen te interpreteren op basis van de canonical coefficients en de intraset correlaties. De uiteindelijke regressie-coefficiënten van de multipele regressie tussen de site scores en de omgevingsfactoren worden de canonical coefficients genoemd. De intraset correlaties zijn de correlatiecoefficienten tussen de omgevingsfactoren en de omgevings assen ttz de ordinatieassen die een lineaire combinatie zijn van de omgevingsfactoren (ENVI AX1 etc. op de listings). Het teken en de relatieve waarden van de intraset correlaties en de canonical coefficients geven informatie over het belang van de diverse omgevingsfactoren bij het voorspellen van de soortensamenstelling. Beide getallen geven dezelfde informatie in het bijzondere geval dat de omgevingsfactoren onderling niet gecorreleerd zijn. Bij veldgegevens is dit meestal niet het geval en dan geven beide grootheden verschillende informatie. Zowel de canonical coefficients als de intraset correlaties zijn gerelateerd aan de snelheid waarmee de beschouwde gemeenschap verandert per eenheid van verandering in de desbetreffende omgevingsfactor maar bij de canonische coefficiënt is het zo dat het effect van de andere omgevingsfactoren constant gehouden wordt, terwijl bij de intraset correlaties er aangenomen wordt dat de overige omgevingsfactoren meevariëren op dezelfde manier als in de dataset. Er dient hier een belangrijke opmerking gemaakt te worden. Wanneer een aantal omgevingsvariabelen sterk onderling gecorreleerd zijn dan treedt het probleem van "multicollineariteit" op. Het effect van de verschillende omgevingsvariabelen op de gemeenschap kan niet ontrafeld worden. In dit geval zijn de canonical coefficients onstabiel en kunnen ze niet geïnterpreteerd worden. De intraset correlaties zijn daar niet aan onderhevig en kunnen verder gebruikt worden voor de interpretatie. Het valt in zo'n geval evenwel aan te raden om een aantal omgevingsfactoren uit de analyse weg te laten. De eigenwaarden en de species-environment correlaties zullen slechts weinig dalen. Indien ze dat toch doen hebben we de verkeerde factoren verwijderd. De multipele correlatie coefficient van de regressie tussen de sitescores en de omgevingsfactoren worden de species-environment correlation genoemd. Dit is dus de correlatie tussen de sitescores (gewogen gemiddelden van de soort-scores) en de sitescores die een lineaire combinatie zijn van de omgevingsvariabelen. Het is een maat van de associatie tussen soorten en omgeving maar de eigenwaarde van de assen zijn eigenlijk een betere maat gezien assen met een kleine eigenwaarde misleidend hoge species-environment correlatie kunnen hebben. De eigenwaarde meet effectief hoeveel van de variatie in de soortgegevens die verklaart wordt door de as en bijgevolg gevolg ook door de omgevingsfactoren.

Naast deze analyse is het vooral de visuele weergave van het ordinatiediagram of de soorts-omgevings biplot die de kracht van CANOCO uitmaakt (Fig. 6.23). De soorten en monsters kunnen in een twee (of drie) dimensionele ruimte weergegeven worden. Gezien de positie van elk monsters in het ordinatie vlak in het centroid ligt van de soortspunten die in dit monster voorkomen, kunnen we uit de figuur afleiden welke soorten bij welke monsters horen. De abundantie of de kans van voorkomen van een soort neemt af met de afstand tot zijn locatie in de figuur.

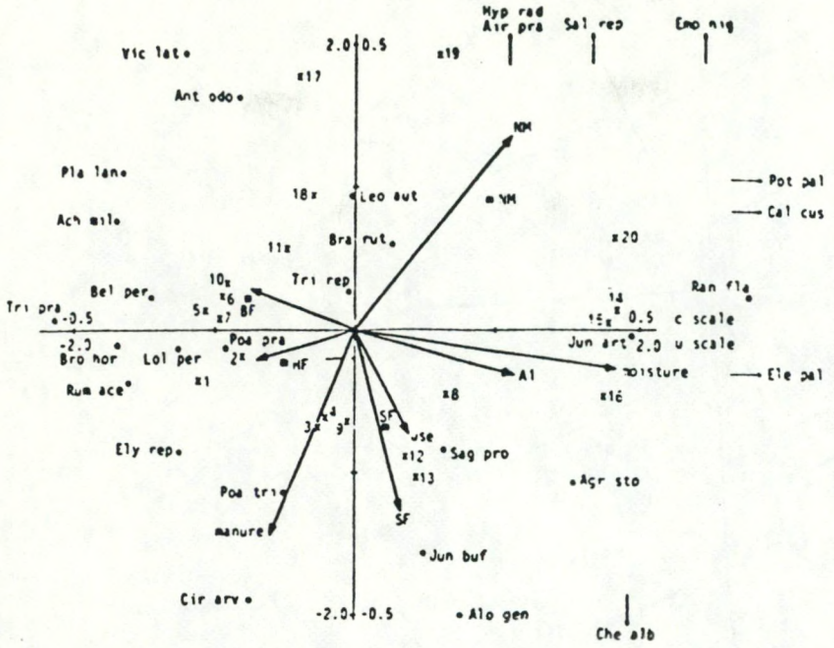


Fig. 6.23 voorbeeld van een CCA met soort en staalpunten en vectoren.

In additie tot het weergeven van de soorten en monsters in de figuur kunnen we ook de omgevingsvariabelen toevoegen en wel op de volgende manier: elke omgevingsvariabele is een vector (pijl), waarvan de positie afhangt van de eigenwaarde van de as en de intraset correlaties. De coördinaten van de top van de pijl op as s zijn:

$$Rjs * \sqrt{lambda_s * (1 - lambda_s)}$$

met Rjs de intraset-correlatie van omgevingsfactor j en lambda de eigenwaarde van as s. Omgevingsfactoren met lange pijlen zijn meer gecorreleerd met de assen dan deze met korte pijlen en zijn bijgevolg meer verbonden aan het patroon van variatie in de soortensamenstelling zoals is weergegeven in het ordinatiediagram. Enkel de richting en de relatieve lengte van de pijlen is zinvol, waardoor hun lengte kan aangepast worden om zo goed mogelijk in het ordinatiediagram te passen. Vanzelfsprekend moeten de onderlinge verhoudingen bewaard blijven. De interpretatie van dit diagram gaat als volgt. Elke pijl stelt een as voor in het diagram en de soortpunten moeten loodrecht daarop geprojecteerd worden (Fig. 6.24). De volgorde van de projectie komt nu grosso modo overeen met de ranking van de gewogen gemiddelden van de soorten met respect tot deze omgevingsfactor. Het geeft met andere woorden de positie van de soortcurve weer langs de omgevingsgradiënt.

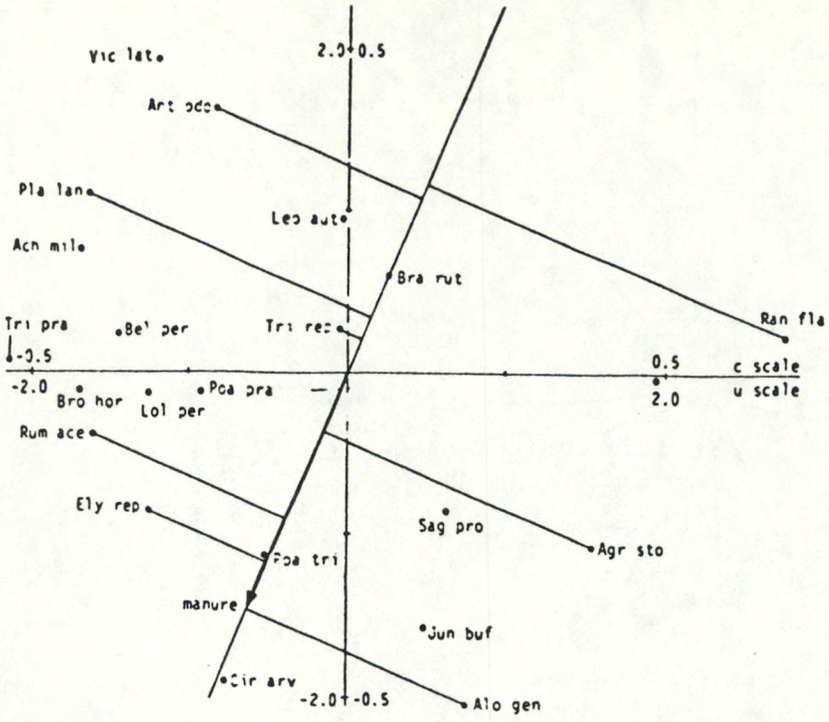


Fig. 6.24. Voorstelling van een ordinatiediagram waarin de ligging van de soorten ten opzichte van een omgevingsvariabele is weergegeven door een loodrechte projectie op de assen.

Bij de keuze van omgevingsfactoren kunnen ook nominale waarden gebruikt worden. Deze worden dan als zogenoemde "dummy variables" in de multipele regressie ingevoerd. Wanneer we bijvoorbeeld het sediment gekarakteriseerd hebben in drie klassen (turf, zand en klei) dan kunnen we dit in de analyse inbrengen als 2 variabelen: turf en zand. Gaat het om een turfbodem dan krijgt die variabele de waarde 1 en zand 0, voor een zandbodem juist andersom en voor een kleibodem krijgen ze allebei 0. Deze nominale variabelen kunnen ook als pijlen voorgesteld worden en de projectie van een soort op dergelijke pijl is een benadering van de totale abundantie van die soort die bereikt wordt in monsters van die klasse. Het is evenwel vaak wenselijker om elke klasse van zo'n nominale variabele voor te stellen als een punt bij de centroid (het gewogen gemiddelde) van de monsters uit die klasse.

VERDERE UITWERKING VAN CANOCO

Het programma CANOCO biedt echter veel meer dan wat we hierboven beschreven hebben. Wat Pielou (1984) aanhaalt voor TWINSPAN is zeker waar voor CANOCO: "Whenever a simple basic method of analysis is refined and elaborated, the number of possible modified forms of the original method increases exponentially and choosing among them becomes increasingly subjective".
Vooreerst kan het programma PCA, RA en DCA, als methoden van "indirect gradient analysis" uitvoeren. We kunnen hier nog opmerken dat PCA een lineair responsmodel en RA en DCA een unimodaal responsmodel veronderstellen. Dezelfde assumpties gelden verder eveneens. Als "direct gradient analysis" hebben we de keuze tussen RDA (Redundancy analysis), CCA (Canonical correspondence analysis, zoals hierboven besproken) en DCCA (Detrended Canonical correspondance analysis). Het verschil tussen CCA en DCCA is in principe hetzelfde als tussen RA en DCA. Het enige bijkomende is dat we nu bij DCCA (en ook bij DCA) detrending niet alleen kunnen uitvoeren op basis van segmenten zoals in DECORANA maar ook via 2de, 3de of 4de orde polynomialen. Detrending op basis van polynomialen zou meer stabiel zijn. Gezien het boogeffect op de

tweede RA-as meestal een kwadratische functie is (een 2de orde polynomiaal) van de eerste as kan detrending met zo'n polynomiaal dit effect extra verwijderen. De hogere polynomiaal functies zouden dan ook kubische relaties tussen de eerste en de derde en kwartische (?) relaties met de vierde as wegwerken. Ter Braak (1986) raadt dan ook sterk aan om deze methode te gebruiken.
Naast de "direct and indirect gradient analysis" biedt het programma ook nog een drietal hybride technieken aan. Hierin kan men bepalen hoeveel assen canonisch zullen zijn of m.a.w. hoeveel assen de beperking krijgen dat ze een lineaire combinatie zijn van de omgevingsfactoren. De overige assen vertegenwoordigen dan de residuele variatie in de data na de extractie van de vorige assen en worden daarom "partiële" assen genoemd. Daarbovenop is er nog een tiende optie voor "nonstandard analysis" waar de gebruiker ongewone opties of ongewone combinaties van opties kan invoeren.

Het programma voorziet ons verder nog van een heleboel extra informatie waarop we nu kort even willen ingaan.

De interset correlaties zijn de correlaties tussen de omgevingsvariabelen en de sitescores die afgeleid zijn uit de soortgegevens.

Het kan aangetoond worden dat CCA een benadering is van de maximum likelihood oplossing van de Gaussiaanse ordinatie wanneer aan bepaalde voorwaarden voldaan is (Ter Braak, 1986b, 1987b). Toch blijkt CCA vrij robust wanneer aan deze voorwaarden niet voldaan is. Essentieel is echter dat we uitgaan van unimodale responscurven. In het geval van simpele monotone relaties kunnen we verwachten dat de resultaten nog steeds voldoen maar bij meer complexe modellen is de methode niet meer geldig.

covariabelen

M.b.v. CANOCO kan ook statistisch nagegaan worden welke soorten gekorreleerd zijn met de gespecifieerde variabelen. Enerzijds bestaat de mogelijkheid om de ordinatiescores te korreleren met de milieuvariabelen (de "indirecte" ordinatie benadering). Een meer gesofistikeerde methode vormt de Monte Carlo permutatie test (Hope, 1968). De waarden van de milieuvariabelen worden op een willekeurige wijze herverdeeld over de verschillende monsters. Vertrekkend van de nulhypothese dat de soorten ongekorreleerd zijn met de milieuvariabelen, mag de oorspronkelijke eigenwaarde niet binnen de 5% hoogste eigenwaarden liggen. Is dat wel het geval dan kan men stellen dat de beschouwde milieuvariabele(n) wel degelijk een invloed op de ordinatie uitoefenen en dat de soorten wel gekorreleerd zijn met de beschouwde variabelen. Het effect van een bepaalde variabele kan nagegaan worden door eliminatie van de eventuele effecten van andere variabelen, door deze laatste te specifiëren als covariabelen.

Voor een meer gedetailleerde beschrijving van deze techniek verwijzen we naar Ter Braak (1985, 1986 en 1987) en Fangstrom et al. (1987).

7) KLASSIFICATIE OF CLUSTERANALYSE

Klassificatie is het toekennen van entiteiten (monsters, kwadraten, taxa, individuen, ...) tot groepen of klassen op basis van hun gelijkenis.

De noodzaak om alles in kategorieën op te delen is volgens Simpson (1961 in Clifford & Stephenson, 1975) een algemeen kenmerk voor alle levende organismen. Amoeben kunnen, net als mensen, niet voortbestaan zonder de kategorie "voedsel" te onderkennen.
Zoals Clifford & Stephenson (1975) het stellen : "Human thought in general, as reflected in human language, seems greatly dependent on the recognition of groups ...". De grootte van eender welke dataset -hetzij bvb. visuele gegevens die in ons brein moeten geassocieerd worden met, aan de waargenomen voorwerpen verbonden karakteristieken, hetzij de resultaten van een wetenschappelijk onderzoek met een of andere specifieke vraagstelling - wordt snel een limiterende factor bij de diepgang van de analyse. Daarover zegt Savory (1970 in Clifford & Stephenson, 1975) - in de kontext van taxonomische klassifikatie - het volgende : " We find so many different animals in the world that we cannot treat them separately and even if we wanted to do so, the task would be beyond the capacity of the human mind and memory. Classification is forced upon us by the limitations of the brain.".

Klassificatie kan voor diverse doelstellingen gebruikt worden (Greig-Smith, 1980): 1) klassificatie als doelstelling op zich of als basis voor inventarisatie of kartering; 2) klassificatie als identificatie van echte entiteiten zoals levensgemeenschappen; 3) klassificatie als middel bij de exploratie van de correlaties tussen vegetatie en milieu.

We kunnen een onderscheid maken tussen formele en informele methodes. Formele methodes zijn reproduceerbaar en dit i.t.t. informele technieken (zoals bijvoorbeeld de Braun- Blanquet methode uit de plantkunde), die steunen op de ervaring en intuïtie van de onderzoeker en daarom verschillende resultaten kunnen gven bij herhaling. Hier beperken we ons tot het bespreken van formele methodes.

De meeste klassificatietechnieken bestaan uit twee verschillende fasen. Vooreerst gaan we tussen alle entiteiten (hetzij monsters, hetzij soorten) een bepaalde (dis)similariteit berekenen op basis van een aan de data aangepaste index, en daarna gaan we de entiteiten op basis van deze similariteiten groeperen. Er zijn ondertussen enorm veel indices uitgewerkt, elk met specifieke eigenschappen. De keuze van de index zal in grote mate de resultaten bepalen vandaar dat we hierop evèn dieper zullen ingaan.

SIMILARITEITS-INDICES

Een similariteitsindex zet, zoals gezegd, de data matrix, bestaande uit de karakterisieken van elke entiteit, om in een "similariteiten"-matrix. Hierin is tussen elke entiteit de gelijkenis in de vorm van een getal weergegeven. Er bestaan zowel similariteitsindices als dissimilariteitsindices. Laatstgenoemde kennen lage waarden toe aan gelijke monstereenheden, eerstgenoemde zeer hoge.

Monstereenheden of entiteiten kunnen zowel kwalitatief, op basis

van de aan- of afwezigheid van de kenmerken (binaire data), als kwantitatief, op basis van gemeten waarden van de kenmerken (bv. dichtheden of biomassa's) (continue data) vergeleken worden. Verder kunnen (dis) similariteitsindices metrisch zijn of niet. Een index is metrisch wanneer het aan bepaalde voorwaarden voldoet (Clifford & Stephenson, 1975):

1) Symmetrie.
gegeven 2 entiteiten (x,y) dan geldt dat de afstand d voldoet aan:

$$d(x,y) = d(y,x) >= 0.$$

2) "Triangular inequality".
gegeven drie entiteiten (x,y,z) dan geldt dat de afstanden tussen hen, d(x,y), d(x,z), d(y,z) voldoen aan:

$$d(x,z) <= d(x,y) + d(x,z).$$

3) "Distinguishability of nonidenticals".
gegeven twee entiteiten (x,y):

if d(x,y) ne 0, then x ne y. (ne = not equal)

4) "Indistinguishability of identicals".
gegeven twee identieke entiteiten (x,x) dan geldt dat de afstand d:

$$d(x,x) = 0.$$

Volgens Clifford en Stephenson (1975) zijn metrische indices te verkiezen omdat ze dezelfde geometrische eigenschappen hebben als de euclidische afstand.

Er bestaan verwarrend veel verschillende indices, waarvan we er hier slechts een aantal - weliswaar de meest succesvolle en meest frequent gebruikte - voorstellen (voor een meer volledig overzicht zie Clifford & Stephenson, 1975). Vergelijkende studies tussen verschillende indices zijn terug te vinden in, Campbell (1978), Wolda (1981), Bloom (1981), Faith (1987) en Develter (1985). Een voorbeeld, waarbij de diverse indices werden uitgerekend voor een hypothetische dataset, is samengevat in Tabel 7.1. Bij de bespreking van enkele eigenschappen van de indices wordt naar deze voorbeelden verwezen.

Referenties????

SORENSEN

Deze kwalitatieve similariteitsindex baseert zich op soortensamenstelling en wordt dus uiteraard niet beïnvloed door transformaties. Normaal gezien wordt de dataset best ook niet gereduceerd, tenzij het zeer zeldzame soorten betreft. Deze index is absoluut onbruikbaar indien het vrij homogene gemeenschappen betreft.

$$S = 2 * c / (a + b)$$

met a en b respectievelijk het aantal soorten in monster 1 en 2 en c het aantal gemeenschappelijke soorten.

RENKONEN INDEX OF PERCENTAGE SIMILARITEIT

Deze kwantitatieve index wordt sterk beïnvloed door gelijke verhoudingen (zoals in vb. 1) en door grote waarden (vb. 5) en reageert slechts zwak op gelijke aantallen (vb. 4).

$$R = \sum_{i=1}^{s} (\mathrm{Min}(P_{1i}, P_{2i}))$$

met $P_{ji} = X_{ji} / N_j$ waarbij X het aantal individuen van soort i en N het totaal aantal individuen in monster j voorstelt en s is het totaal aantal soorten.

BRAY CURTIS INDEX

Deze index verandert sterk met de monstergrootte (Wolda, 1981), hetgeen geen probleem vormt indien het enkel de bedoeling is monsters van gelijke grootte met elkaar te vergelijken. Deze index wordt evenwel zeer sterk beïnvloed door grote aantallen (uitschieters), hetgeen eventueel kan opgevangen worden door logaritmeren (vb. 5). Gelijke verhoudingen leiden niet tot een hoge similariteit (vb. 1) en gelijke aantallen van eenzelfde soort verhogen de similariteit evenmin (vb 4). Net als de Canberra metric is dit in feite een dissimilariteitindex. De similariteit wordt dan uiteraard bekomen door de waarde van 1 af te trekken. Het is volgens Bloom (1981) de enige index die nauwkeurig similariteiten weergeeft, zonder hoge of lage similariteiten te benadrukken.

$$BC = 1 - \frac{\sum_{i}^{s} :(X_{1i} - X_{2i}):}{\sum_{i}^{s} (X_{1i} + X_{2i})}$$

met X_{1i} en X_{2i} de aantallen in monster 1 , respectievelijk monster 2 van soort i en s het totaal aantal soorten.

CANBERRA METRIC INDEX

Ook deze index wordt sterk beïnvloed door de monstergrootte, en stijgt bovendien niet lineair van 0 tot 1 (Wolda, 1981). Canberra overschat lage en onderschat hoge waarden, zodat uiteindelijk veel knooppunten in het midden van het dendrogram liggen (Bloom, 1981). Omdat deze index het gemiddelde is van een reeks breuken, zodat een groot getal slechts tot één van de fracties bijdraagt, is Canberra minder gevoelig voor dominantie (vb. 5), wat voor sommige datasets een groot voordeel kan zijn. Canberra en Bray Curtis reageren het best op gelijke aantallen (vb. 4) en zijn ongeveer even ongevoelig voor gelijke verhoudingen (vb. 1).

$$C = 1 - \sum_{i=1}^{s} \frac{:(X_{1i} - X_{2i}):}{(X_{1i} + X_{2i})}$$

met X_{1i} en X_{2i} de aantallen in monster 1 , respectievelijk monster 2 van soort i en s het totaal aantal soorten.

CEKANOVSKI

$$\frac{2 \sum_{i=1}^{n} \min (X_{ij}, X_{ik})}{\sum_{i=1}^{n} (X_{ij} + X_{ik})} 100$$

met X_{ij} de score van soort i in staal j; X_{ik} de score van soort i in staal k en n het aantal stalen.

SIMILARITY RATIO

Deze index werd zeer veel gebruikt in het Nederlands vegetatiekundig onderzoek en is ook beschikbaar in CLUSTAN.

$$\frac{\sum_{i=1}^{n} (X_{ij}, X_{ik})}{\sum X^2_{ij} - \sum X_{ij} X_{ik} + \sum X^2_{ik}} 100$$

EUCLIDISCHE AFSTAND

$$\sqrt{\sum_{i=1}^{n} (X_{ij} - X_{ik})^2}$$

ENKELE KENMERKEN VAN SOMMIGE INDICES

In Tabel 7.1 zijn de waarden van verschillende indices voor een hypothetische dataset weergegeven.

		MONSTERS									
		1		2		3		4		5	
		A	B	A	B	A	B	A	B	A	B
SOORT	1	1	10	30	30	30	30	30	30	30	30
	2	2	20	5	15	5	15	5	15	5	15
	3	3	30	2	5	2	5	2	5	2	10000
	4	4	40	18	24	18	24	18	24	18	24
	5	5	50	20	18	20	18	20	18	20	18
	6	6	60	3	5	3	5	3	5	3	5
	7					0	0	20	20		
	8					0	0	20	20		
	9					0	0	20	20		
	10					0	0	20	20		
SORENSEN		1		1		1		1		1	
RENKONEN		1		0.854		0.854		0.900		0.035	
CANBERRA		0.182		0.771		0.771		0.863		0.676	
BRAY-CURTIS		0.181		0.869		0.869		0.931		0.015	
CANBERRA*		0.430		0.850		0.850		0.910		0.770	
BRAY-CURTIS*		0.490		0.900		0.900		0.940		0.710	

Tabel. 7.1. Overzicht van de similariteiten, berekend volgens de vier aangegeven indices, tussen 5 maal 2 hypothetische monsters. (* na logaritmeren van de data). Van zowel de Bray-Curtis als de Canberra index werd de bekomen dissimilariteit omgezet in een similariteit (1-de bekomen index waarde).

Merken we op dat van deze vier indices er geen enkele is die gemeenschappelijke afwezigheden in rekening brengt - in Clifford en Stephenson, 1975 wordt echter een overzicht gegeven van enkele technieken (o.a. de Sokal en Sneath index) die dat wel doen. Alle vier zijn ze voor een bepaald doel veel efficiënter dan de andere (ook de meeste niet besproken) indices : Sorensen als kwalitatieve index, Renkonen die gevoelig is voor gelijke verhoudingen, Bray Curtis voor absolute aantallen en Canberra als een dominantie-ongevoelige variant van Bray Curtis. De keuze hangt dus integraal af van de aard van de data en van de beoogde doelstellingen.

Er bestaan veel programmapaketten, waarin verschillende similariteitsindices zijn voorzien : CLUSTAN is wel het best bruikbaar, omwille van de uitgebreide keuzemogelijkheden en omdat ook een aantal sorteringstechnieken (zie verder) in dit pakket zijn opgenomen.

SORTERINGSMETHODEN

Een clusterstrategie bestaat er nu in om op basis van de (dis)similariteitenmatrix, bekomen via één of andere similariteits index, die monsters met gelijkende similariteiten in groepen samen te brengen. Dit is in Fig. 7.1 samengevat. Merk op dat de similariteiten matrix slechts voor de helft gevuld is. De andere helft is uiteraard het spiegelbeeld. Ook de diagonaal van de matrix bevat geen informatie, gezien het de similariteit tussen hetzelfde staal betreft. Het omzetten van een similariteitsmatrix in een dendrogram is gebaseerd op een sorteringstechniek. Dit zullen we hier verder behandelen.

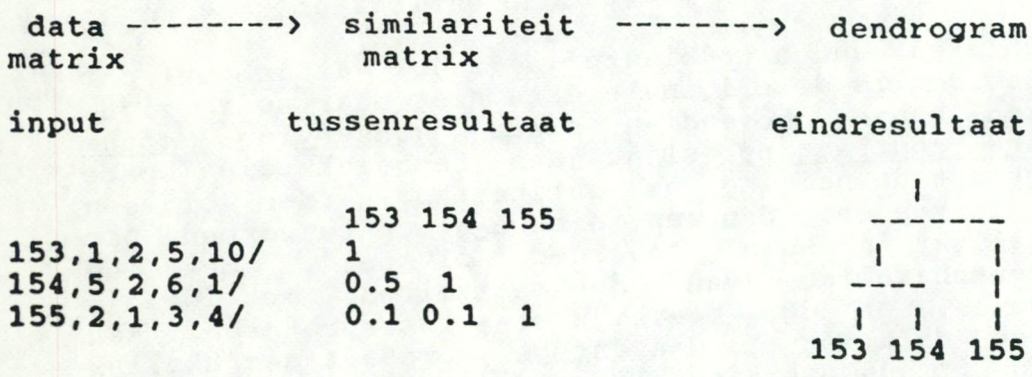


Fig. 7.1 Overzicht van de verschillende stappen bij een clusteranalyse.

Referenties

Niet-hiërarchische klassificatietechnieken delen de entiteiten op in groepen. Hiërarchische technieken daarentegen voegen daar nog een extra dimensie aan toe, namelijk de (dis)similariteit tussen de opeenvolgende groepen, zodat we uiteindelijk tot een soort stamboom, een dendrogram komen.

Divisieve technieken delen successief de dataset in steeds kleinere groepen op, terwijl agglomeratieve technieken stapsgewijs, entiteit per entiteit vergelijken met reeds gevormde clusters en vervolgens de entiteit toevoegen aan de meest gelijkende groep.

Dit toekennen aan een bepaalde groep kan gebeuren op basis van een enkel criterium (monothetisch), of op basis van meerdere criteria (polythetisch), bijvoorbeeld rekening houdend met de dichtheden van meerdere of van alle soorten.

In deze paragraaf zullen we ons beperken tot formele, agglomeratief, polythetische hierarchische clustermethoden. Als enig voorbeeld van een divisieve klassificatiemethode bespreken we in een aparte paragraaf de TWINSPAN techniek.

Bij het sorteren kan op verschillende manieren te werk gegaan worden, al naargelang de criteria die gehanteerd worden voor het berekenen van de similariteitswaarden van een nieuw gevormde cluster t.o.v. de andere groepen.

```
                    similariteit a
entiteit X           <---------->     andere groepen en nog niet
                                        toegekende entiteiten
   +
                    similariteit b
groep Y              <---------->     andere groepen en nog niet
                                        toegekende entiteiten
   =
                    similariteit ??
groep Y' = X + Y     <---------->     andere groepen en nog niet
                                        toegekende entiteiten
```

Sorteringstechnieken kunnen combinatorial zijn, wat inhoudt dat eens een cluster gevormd is de originele data niet meer nodig zijn. Sorteringstechnieken kunnen bovendien space conserving, space dilating of space contracting zijn. Een techniek die space contracting is vertoont de neiging om entiteiten eerder bij een bestaande cluster te voegen, dan een nieuwe groep te vormen. Dit resulteert gemakkelijk in een "chained cluster", met een sequentiële aaneenschakeling van de entiteiten, met weinig onderscheiden groepen, en dus moeilijk te interpreteren. Een techniek die space dilating is daarentegen, creëert gemakkelijk nieuwe groepen en geeft dus overzichtelijke dendrogrammen.
Volgens Lance en Williams (1966 en 1967 in Greig Smith, 1983) zijn "combinatorial" technieken te verkiezen. Hoewel dit ontegensprekelijk een groot voordeel is bij het uitwerken van een sorteringsalgoritme en zeker bij het besparen van computertijd, hoeven ze daarom niet noodzakelijk als beter te worden beschouwd. Bovendien lijkt het concept dat aan de basis ligt van "non combinatorial" technieken, namelijk rekening houden met de heterogeniteit van een groep, intuïtief beter te verantwoorden.

Hieronder worden enkele sorteringstechnieken besproken. In Fig. 7.2 is een overzicht gegeven van die verschillende methoden toegepast op eenzelfde dataset.

NEAREST NEIGHBOUR SORTING (NEN)

De afstand tussen twee groepen wordt gedefiniëerd als de kortst mogelijke afstand tussen twee monstereenheden, één van elke cluster. Deze techniek is space contracting en levert dikwijls chained clusters, die al snel erg ongelijkende monsters bijeen groeperen (Gauch,1982). Deze techniek is non combinatorial (aangezien alle entiteiten met elkaar moeten vergeleken worden). Volgens Pritchard en Anderson (1971 in Clifford et al.1975) is NEN weinig bruikbaar en ongeschikt voor gemeenschapsecologische studies (Hill,1977 in GAUCH,1982).

FURTHEST NEIGHBOUR SORTING (FUN)

Hier wordt de grootste afstand tussen twee monstereenheden gebruikt als maat voor de dissimilariteit tussen twee groepen. Dit is eveneens een non combinatoriale techniek. Zowel NEN als FUN zijn volgens Pielou (1984) weinig bruikbaar.

GROUP AVERAGE SORTING (GAV)

Omdat als maat voor de afstand tussen twee groepen de gemiddelde afstand tussen alle entiteiten van een bepaalde groep en alle entiteiten van de andere groep wordt gebruikt, is deze non combinatoriale techniek space conserving. GAV is een veel gebruikte methode en lijkt volgens Field et al. (1982) de meest succesvolle sorteringstechniek. Ook Sneath en Sokal (1973 in Gauch, 1982) raden GAV aan en dan meer bepaald de "Unweighted Pair Groups Method (using aritmetic) Averages" (U.P.G.M.A.). Deze methode vertrekt steeds van de originele, ongewogen (dis)similariteits indices om de gemiddelde (dis)similariteit bij fusie te berekenen. Volgens deze auteurs is de correlatie tussen de oorspronkelijke similariteitenmatrix en de gesorteerde matrix maximaal met UPGMA. Weighted Average is een gemodifiëerde versie die aan elke groep een gelijk gewicht geeft.

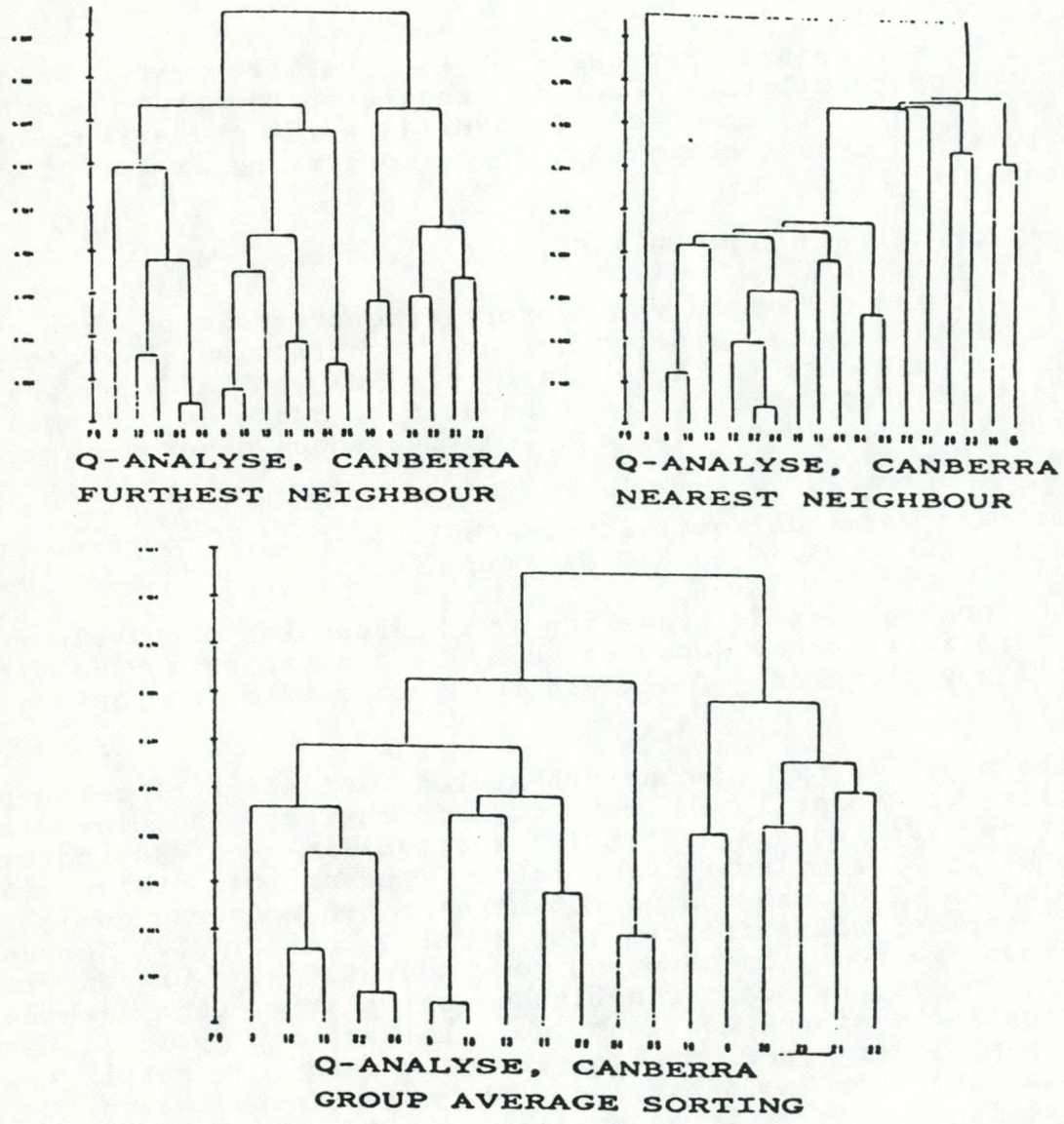


Fig. 7.2 Drie dendrogrammen gebaseerd op drie verschillende sorteringstechnieken toegepast op eenzelfde similariteitenmatrix.

WARDS ANALYSE

Voor de wiskundige achtergrond en de formule wordt verwezen naar Steinhausen en Langer (1977 in Looman,1984), omdat ze nogal complex is. In grote lijnen komt het erop neer dat de variantie binnen de clusters zo klein mogelijk gehouden wordt. Voor elke combinatie van clusters wordt berekend hoeveel de "binnen-cluster-variantie" toe zou nemen als ze bij elkaar gevoegd zouden worden. Er wordt dan gekozen voor de kleinste toename van die variantie. Dit is een populaire methode ondermeer omdat hij meestal clusters levert van een ongeveer gelijk aantal elementen. Als er veel entiteiten te klassificeren zijn, deelt deze techniek soms echter verder op dan ekologisch zinvol is. De vervorming van de similariteiten is bij deze methode veel groter dan bij bijvoorbeeld GAV. Het verdient dan ook aanbeveling deze techniek in combinatie met andere te gebruiken.

FLEXIBLE SORTING (FLEX)

Dit is een sorteringstechniek die kan variëren van space contracting tot space dilating door de zogeheten B-waarde aan te passen (resp. B=+/- 1 en B=0). Gewoonlijk wordt echter met B=-0.25 gewerkt, zodat deze techniek dan space conserving is (Clifford en Stephenson, 1975).

$$d_{hk} = A_i * d_{hi} + A_j * d_{hj} + B * d_{ij} + C * abs(d_{hi} - d_{hj})$$

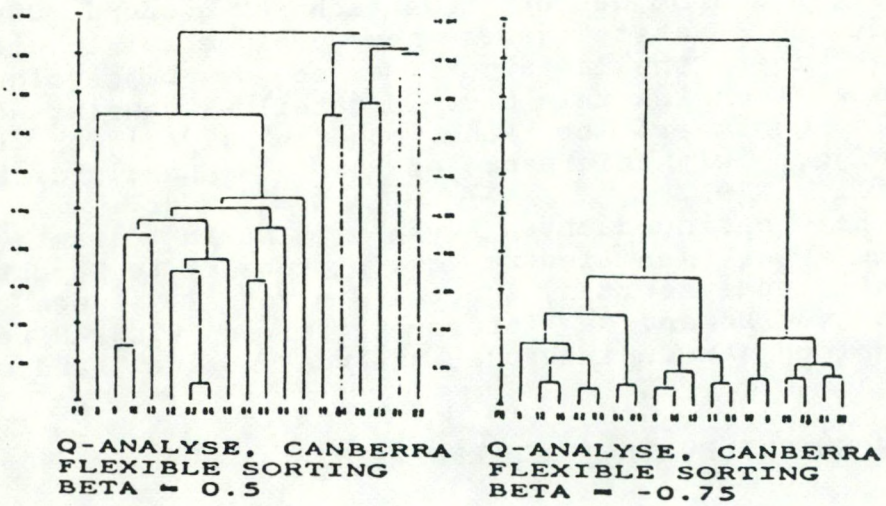
waarbij A_i, A_j, en C de aard van de sorteringstrategie bepalen en waarbij d de afstand tussen groepen h,i en j voorstelt. Uit eigen onderzoek bleek dat deze techniek, net als FUN en GAV trouwens, overzichtelijke dendrogrammen gaf met B=-0.25 en dus goed bruikbaar was. Het effect van Beta op het resulterende dendrogram is weergegeven in Fig. 7.3.

Alle hierboven vermelde sorteringstechnieken zijn voorzien in het programmapakket CLUSTAN en PATIMA (Wageningen).

Merken we nog op dat de indeling in klassen kan gecorreleerd worden met de abiotische factoren d.m.v. een variantie-analyse (continue milieuvariabelen) of met een chi kwadraat toets (nominale variabelen).

Ook is het mogelijk om ordening van stalen (of soorten) bekomen in ordinaties te vergelijken met classificaties. In hoeverre weerspiegelt de ordening van stalen (of soorten) in een dendrogram de oorspronkelijke similariteiten latrix tussen de stalen; in hoeverre gelijken 2 dendrogrammen op elkaar; in hoeverre gelijkt een ordinatie op een classificatie etc. Sokal & Rohlf (1962) hebben een objectieve methode beschreven om dergelijke vragen te beantwoorden, de zogenaamde cofenetische correlatie. Deze methode kent een cofentische waarde toe aan de similariteit volgens het dendrogram van ieder paar stalen en genereert een matrix van dergelijke waarden voor alle stalen. De cofentische waarde C_{ij} tussen 2 stalen j en k is de maximale similariteit volgens het dendrogram tussen de 2 stalen (ook de rangorde van de fusieniveaus is bruikbaar). Als men dit voor alle stalen uitvoert bekomt men een nieuwe matrix van zogenaamde cofentische waarden. Cofentische waarden gebaseerd op ordinaties zijn de aktuele afstanden in de

ordinatie. Finaal wordt de produkt-moment correlatie coëfficient berekend tussen de overeenkomstige elementen van de originele similariteitenmatrix en de matrix met cofentische waarden.



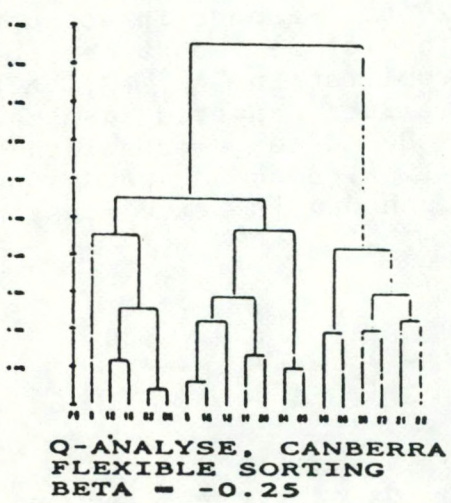


Fig. 7.3 Dendrogrammen van eenzelfde dataset gebaseerd op Canberra metric similariteits index en Flexible sorting maar met drie verschillende waarden van Beta.

EEN HYBRIDE KLASSIFICATIE TECHNIEK : TWINSPAN

TWINSPAN (Two Way INdicator SPecies ANalysis) (Hill, 1979) is een divisieve polythetische clustermethode, die echter gebaseerd is op ordinaties, en daarom een hybriede techniek genoemd wordt. Deze methode wordt de laatste jaren zeer veel gebruikt, niet in het minst door het zeer snelle algoritme en de bijgevolg beperkte rekentijd. Deze techniek zal hier in detail uitgewerkt worden. De tekst is deels gebaseerd op Van Tongeren (1987) en voor meer details verwijzen wij hiernaar of naar de handleiding van het programma (Hill, 1979).

Eén van de basisgedachten van TWINSPAN stamt uit de Fytosociologie waar men iedere groep probeert te karakteriseren door differentiërende soorten, soorten die vooral in één deel van de dichotomie voorkomen. De interpretatie van TWINSPAN resultaten is dan ook analoog aan de interpretatie van met de hand geordende tabellen.

PSEUDOSOORTEN EN CUTLEVELS (DREMPELNIVEAUS)

Gezien de gedachte van differentiërende soorten in essentie kwalitatief is, maar er uiteraard ook met kwantitatieve data moet kunnen gewerkt worden ontwikkelde Hill et al. (1975) een kwantitatief equivalent, de zogenoemde "pseudo-species" of schijnsoort. De hoeveelheid van elke soort wordt vervangen door het voorkomen van 1 of meer pseudosoorten. Hoe abundanter een soort hoe meer pseudosoorten er gedefiniëerd worden. Elke pseudosoort is gedefiniëerd door de minimum abundantie van de corresponderende soort, het zogenoemde CUT-LEVEL. De in het programma ingebouwde cut-levels zijn: 0 2 5 10 20 50. Wanneer in een opname van soort A 49 individuen aanwezig zijn (of de soort een bedekking van 49 % heeft) dan wordt hij opgesplitst in A1, A2, A3, A4 en A5 (Fig. 7.4). Zijn slechts 6 individuen aanwezig dan zal hij opgesplitst worden in A 1 en A 2. De twee voorbeelden hebben dus twee pseudosoorten gemeenschappelijk. Zodoende wordt onderscheid gemaakt tussen de verschillende dichtheden in beide monsters.

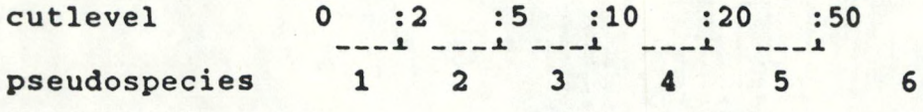


Fig. 7.4 Voorstelling van de cutlevels en pseudospecies zoals gebruikt in het programma TWINSPAN.

Deze manier van het vervangen van een kwantitatieve variabele door meerdere kwalitatieve variabelen wordt "conjoint coding" genoemd (Heiser, 1981). Het grote voordeel daarvan is dat als de soort een unimodale responscurve vertoont langs een gradiënt, elke pseudosoort dit ook doet en als de responscurve geskewd is, de pseudosoorten respons curven verschillen in hun optimum.
De keuze van de cutlevels bepaalt dan ook in belangrijke mate de eigenlijke resultaten en we zullen dan ook verder dieper ingaan op

de keuze van de cutlevels.

HET MAKEN VAN DICHOTOMIEEN

Monsterclassificatie

TWINSPAN maakt geordende twee-wegs (kruis) tabellen aan de hand van differentiërende soorten. Het gelijkt dan ook sterk op de methode van Braun-Blanquet voor het met de hand sorteren van tabellen. In tegenstelling tot deze methode waar soorten en monsters gelijktijdig worden geordend gaat TWISPAN eerst de monsters en pas daarna de soorten ordenen en dit gebaseerd op de monsterklassificatie.

De opslitsing van de data in verschillende groepen komt tot stand door opeenvolgende opsplitsingen die telkenmale gebaseerd zijn op drie opeenvolgende ordinaties. De gebruikte ordinatiemethode is correspondentie analyse of reciprocal averaging. Een eerste ruwe dichotomie wordt gemaakt op basis van een ordinatie van de monsters. De eerste ordinatieas wordt gesplitst in twee stukken ter hoogte van de centroid. De groepen worden de negatieve (linkse) en positieve (rechtse) groep genoemd. In deze stap gebeurt dus de identificatie van de richting van variatie en verkrijgt men een ruwe opdeling van de monsters.

In een tweede stap worden de differentiërende soorten geïdentificeerd, die preferentiëel aan één kant van het centroïd voorkomen. Een preferentie-score van +1 wordt gegeven aan elke soort die drie maal frequenter voorkomt aan de positieve dan aan de negatieve kant, en die algemener is dan een bepaald minimum. Negatief preferentiële soorten krijgen dan uiteraard -1. Zeldzamere soorten of soorten die minder preferentiëel zijn krijgen een lager gewicht of preferentie-score. Dit kunnen we simpel als volgt voorstellen:

laat soort j een frequentie AYY en AY hebben aan de positieve, respectievelijk negatieve zijde dan is:

$$\text{preferentie ratio (PREF)} = (AYY-AY)/(AYY+AY)$$

met $\text{frequentie (FREQ)} = AYY+AY$

als of de frequentie van de soort lager is dan een threshold (20% van de monsters of 0.2) of de preferentie is lager dan een threshold (drie maal frequenter aan de ene zijde dan aan de andere) dan wordt de volgende "downweighting" toegepast:

$$\text{preferentie score} = (FREQ/0.2)*(PREF/3)**5.$$

Hieruit blijkt duidelijk dat zeldzame soorten een lager gewicht krijgen maar dat soorten met een onvoldoende preferentie-ratio een zeer veel lager gewicht krijgen.
Het eerste deel van de tweede ordinatie bestaat er dan in om de preferentie-scores van alle soorten per monster op te tellen en de som daarna te standardiseren zodat de hoogste waarde 1 is.
Het tweede deel bestaat uit het berekenen van de gemiddelde preferentiescores per monster zonder "downweighting" van de zeldzamere soorten. Hierdoor kunnen deze laatste ook nog hun invloed laten gelden. In vergelijking met de eerste ordinatie polariseert deze weinig wanneer er veel algemene soorten zijn (non preferentials), wat te verwachten is op een lager niveau van de

hiërarchie. Op een hoger niveau polariseert het des te meer omdat hier nog vele zeldzame soorten aanwezig zijn.
De uiteindelijke "refined ordination" wordt verkregen als de som van beide ordinaties (de scores per monster worden opgeteld en gesorteerd) en wordt dan gesplitst op een punt in het midden. Met de uitzondering van enkele "borderline cases", bepaalt deze ordinatie de uiteindelijke dichotomie. Deze refined ordination wordt voor technische redenen in 16 segmenten onderverdeeld (Fig. 7.5). Segmenten 5 tot 12 beslaan 20% van de lengte van de totale ordinatie en worden de "critical zone" genoemd. De lengte van de segmenten binnen deze zone beslaat 1/4 van die erbuiten of 2.5% van de lengte van de ordinatie. Binnen deze "critical zone" worden 4 segmenten afgebakend die de "zone of indifference" genoemd worden. Deze zone kan op 5 plaatsen binnen de "critical zone" geplaatst worden. De keuze wordt zo gemaakt dat het aantal "misclassified samples" minimaal is (zie verder). De monsters die in de "zone of indifference" liggen worden afhankelijk van hun plaats borderline positive of negative genoemd.
Voor deze "borderline cases" valt de definitieve beslissing over de groepsindeling in de laatste ordinatie of de "indicator ordination". Eerst worden hiervoor de "indicators" bepaald en wel als volgt: voor alle soorten wordt een preference index opgesteld:

preference index= kans van voorkomen op de + kant min kans van voorkomen op de - kant

Bijvoorbeeld voor een soort die in 30% van de monsters aan de positieve en in 10% van de monsters aan de negatieve zijde voorkomt is de preference index = 0.3-0.1=0.2. Met een ingestelde threshold (FEEBLE) van 0.1 is deze soort dus een indicatorsoort. Hierbij dienen evenwel nog twee opmerkingen gemaakt te worden. Ten eerste wordt de absolute waarde van de preferentie-index gebruikt waardoor zowel indicatoren voor de positieve als voor de negatieve zijde worden gezocht. Ten tweede, en veel belangrijker, is dat er rekening gehouden wordt met monsters die dicht bij de splitsing van de ordinatie (de borderlines in de critical zone) liggen en wel door die een lager gewicht te geven. Deze weging gaat van 0 voor een staal in het centrum van de ordinatie op een lineaire schaal tot 1 voor stalen op de grens van de critical zone (= 40 en 60% van de uiteinden van de ordinatie). Op de output worden de aldus gekozen indicator-soorten weergegeven in volgorde van belangrijkheid. De weging van monsters in de critische zone verklaart dan ook dat een soort die bv. in 12 van de 15 monsters aan de positieve en 0 monsters aan de negatieve zijde een betere indicator kan zijn dan een soort met 14 versus 0 scores. Immers voor deze laatste soort zullen een aantal stalen een lager gewicht gekregen hebben omdat ze dicht bij het centrum van de ordinatie lagen.
De indicator-ordinatie bestaat nu uit het sommeren van de scores van elke indicator-soort per monster. In Fig. 7.5 is een schema weergegeven van een refined en de daarop volgende indicator ordinatie. De opslitsing van deze laatste is zo gekozen dat het aantal "misclassified samples" minimaal is. In Fig. 7.6 is de ligging van positieve, negatieve, borderlines en misclassified samples weergegeven. De borderlines liggen in de "zone of indifference". Misclassified samples liggen links of rechts van de "zone of indifference" maar worden door de indicator-ordinatie in een andere groep geplaatst dan in de refined-ordinatie. Dit wordt

beschouwd als een mislukking van de indicator ordinatie om de dichotomie van de refined-ordination na te bootsen, vandaar dat deze monsters "misclassified (positive or negative)" noemen. De opdeling van de indicator ordinatie bepaalt dus de ligging van de borderlines.

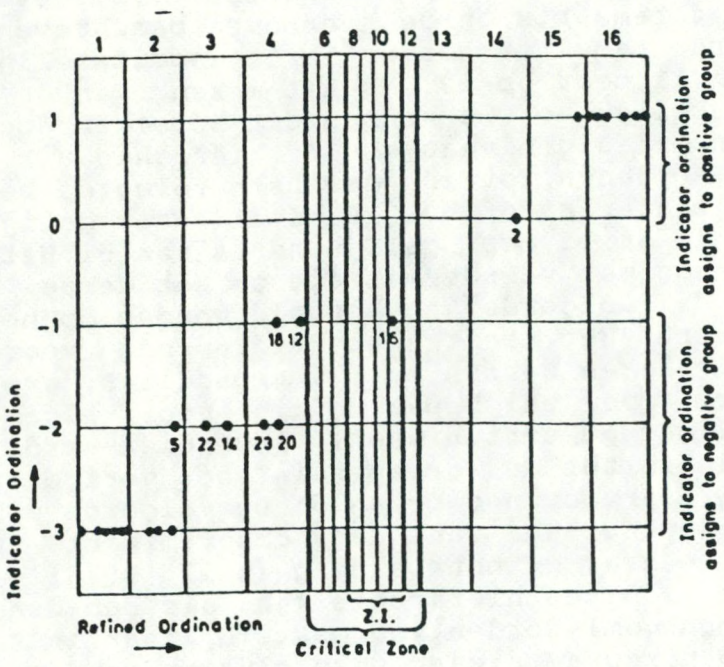


Fig. 7.5. Verband tussen de indicator-ordinatie en de refined-ordinatie. Voor uitleg zie tekst.

```
 1 2 3 4 5 6 7** 8 91011**1213141516****
****************************************
Misclassified **Border- ** Positives** 1
Negatives      **line +ve**          ** 0
****************************************
                **Border- **          **-1
                **line    **Misclassif**-2
Negatives       **negative**Positives **-3
                  Zone of
                  Indifference
 1 2 3 4 5 6 7** 8 91011**1213141516****
****************************************
 0 0 0 0 0 0 0** 0 0 . 1** 0 0 0 1 6** 1
 0 0 0 0 0 0 0** 0 0 0 0** 0 0 1 0 0** 0
****************************************
 0 0 0 2 0 0 0** 0 0 0 1** 0 0 0 0 0**-1
 0 1 2 2 0 0 0** 0 0 0 0** 0 0 0 0 0**-2
 6 3 0 0 0 0 0** 0 0 0 0** 0 0 0 0 0**-3
```

Fig. 7.6. Overzicht van de ligging van de verschillende types monsters in de TWINSPAN classificatie.

Dit ganse splitsingsproces wordt herhaald tot elke cluster niet meer dan een bepaald minimum aantal monstereenheden bevat.

Ordenen van dichotomieën

Beschouw de volgende hiërarchie (Fig. 7.7), gevormd door het opsplitsen van de datamatrix op de hierboven beschreven manier. Om een mooie geordende tabel te hebben is het noodzakelijk dat bv. 10 meer lijkt op 9 en 11 meer op 12. Na het maken van de dichotomieën is dit niet noodzakelijk zo. We kunnen dit bereiken door eerst alle splitsingen op een bepaald niveau van de hiërachie uit te voeren en die daarna zo te ordenen tot de gewenste volgorde bereikt is. In TWINSPAN gaat men echter na of bv. 10 of 11 meer gelijkend is aan 4, of indien bv. 14 of 15 meer gelijkend is aan 6. Het vergelijken van bv. de groepen op het vierde met die op het derde niveau heeft het voordeel dat de volgorde kan bepaald worden op het moment dat de opslitsing gebeurt. Het heeft het bijkomende voordeel dat de nieuwe groepen vergeleken worden met, vermoedelijk, grotere groepen waardoor de volgorde meer zal bepaald zijn door algemene relaties dan door toevallige gebeurtenissen. Bovendien wordt in TWINSPAN niet alleen de relatie tot de groepen in het vorige niveau, maar ook tot het niveau daarvoor nog bepaald. De volgorde zelf wordt dan bepaald door het opstellen van een discriminant functie. Voor details verwijzen we naar de handleiding (Hill, 1979).

De resulterende monster-hiërarchie kan als een dendrogram (een binaire opsplitsingsboom) worden voorgesteld, maar in tegenstelling met een normaal dendrogram, waar elke cluster rond zijn knooppunt mag draaien, moet een TWINSPAN dendrogram als rigied geïnterpreteerd worden, met de meest gelijkende monsters of soorten dichtst bij elkaar (Pielou, 1984).

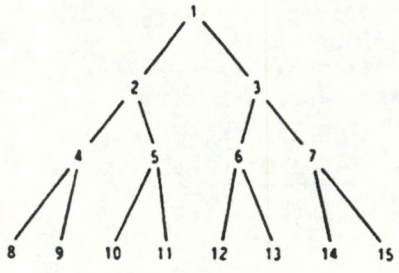


Fig. 7.7. Voorbeeld van een dichotomie

Soortclassificatie

De soorten worden door TWINSPAN op grosso modo dezelfde manier opgesplitst als de monsters. Het grote verschil is evenwel dat het gebeurt op basis van de monster-classificatie en niet op basis van de ruwe gegevens. Ze is gebaseerd op "getrouwheid" (fidelity), dit is de mate waarin een bepaalde soort gebonden is aan een bepaalde groep stalen. De getrouwheid van soort J in de monsters van groep IC is:

$$\text{getrouwheid (RAT(IC,J))} = \frac{\text{gemiddeld voorkomen van J in groep IC}}{\text{gemiddeld voorkomen van J buiten groep IC}}$$

met gemiddeld voorkomen van J als het niet gewogen gemiddelde van de waarde b_{ij} voor de monsters i die behoren tot groep IC en waarvoor b_{ij} = C als de soort J het cutlevel C bereikt in staal i, of maw de gemiddeldes van de het hoogste cutlevel dat elke soort bereikt in de stalen van deze groep. Dit zijn ook de cutlevels die weergegeven zijn in de uiteindelijke two-way tabel op de listing. Bij de classificatie worden dan nog verschillende gewichten toegekend en wel als volgt:

1) extra gewicht voor een hoge getrouwheid;
2) extra gewicht voor algemenere soorten;
3) extra gewicht voor grotere groepen en hogere niveaus in de hiërarchie.

TWO-WAY TABEL

Na het groeperen van monsters en soorten en het ordenen van de groepen produceert TWINSPAN een two-way tabel waarbij voor elke soort het hoogste cutlevel wordt weergegeven (Fig. 7.8). Deze tabel is eigenlijk het belangrijkste resultaat van TWINSPAN. Het grote nadeel bij de interpretatie van deze tabel is dat TWINSPAN binnen elke groep zowel de monsters als de soorten sequentiëel (volgens hun volgnummer) weergeeft, en niet in de volgorde bepaald door de onderliggende ordinatie. De groepen zijn in de juiste volgorde geplaatst, maar binnen de groepen niet de individuele monsters. Dit wordt best opgelost door (voorlopig) manuele herordening. Is de volgorde van groepen rigied, toch kunnen de individuele monsters op het laagste niveau van de hierachie naar eigen inzicht verplaatst worden. Dit kan soms de tabel zeer ten goede komen, vooral wanneer bepaalde groepen uit vrij veel monsters bestaan.

```
 -    1111   1      111112
     17895670123489234560

 3 Air pra .2.3.............. 00000
12 Emp nig ...2.............. 00000
13 Hyp rad 22.5.............. 00000
28 Vac lat 2.1....1.......... 00000
 5 Ant odo .4.44324........... 00001
18 Pla lan 323.5553........... 00010
 1 Ach mil .2..222413......... 000110
26 Tri pra ....252............ 000110
 6 Bel per ..2.2..2.322....... 000111
 7 Bro hor ....2.24.4.3....... 000111
 9 Cir arv ...........2........ 000111
11 Ely rep ....4...4444.6..... 001
17 Lol per 7.2.2666756542..... 001
19 Poa pra 413.2344445444.2.... 001
23 Rum ace ....563.....22..... 001
16 Leo aut 525@3333.522322222.2 01
20 Poa tri ....64542654549..2. 01
27 Tri rep 3.222526.521233261.. 01
29 Bra rut 4.632622..22224..444 01
 4 Alo gen .........2725385..4. 10
24 Sag pro 2..3.......52242.... 10
25 Sal rep ..33.............5 10
 2 Agr sto ..........4843454475 110
10 Ele pal ...........4...4584 11100
21 Pot pal ................22.. 11100
22 Ran fla ...........2..22224 11100
30 Cal cus ................4.33 11100
14 Jun art ...........44...334 11101
 8 Che alb ..............1.... 1111
15 Jun buf ......2......443.... 1111

          00000000000011111111
          00001111111100001111
              00001111
```

Fig. 7.8. Voorbeeld van een Two-Way tabel.

Enkele bedenkingen bij het gebruik van cutlevels

TWINSPAN geeft automatisch meer gewicht aan abundante en algemene soorten d.m.v. het toekennen van verschillende gewichten aan elke pseudospecies. De gewichten zijn namelijk evenredig met de cutlevelklasse waarbinnen de pseudospecies valt (Ter Braak, 1982). Deze default werkwijze kan op verschillende manieren geaccentueerd of juist teniet gedaan worden, zodat we uiteindelijk d.m.v. het gebruik van cutlevels eenzelfde dataset op totaal verschillende manieren kunnen benaderen en analyseren :

* Cutlevels bepalen de eigenlijke dataset waarmee het programma zal werken. Hoe meer cutlevels men specifiëert, hoe meer pseudospecies zullen berekend worden en hoe groter de kwantitatieve differentiatie is (op voorwaarde dan wel dat de cutlevels een grote range bestrijken en vrij uniform verdeeld zijn). In dat geval hecht men m.a.w. veel belang aan de absolute aantallen, waarin de verschillende soorten voorkomen.

* Gebruikt men weinig cutlevels, die bovendien laag liggen, zodat ook de minder abundante (dichtheid) of minder productieve (biomassa analyse) soorten geregeld deze cutlevels bereiken, dan wordt bij de analyse minder gewicht (letterlijk en figuurlijk) gegeven aan de absolute aantallen en worden bepaalde verschillen tussen de soorten onderling geminimaliseerd. Zo worden soorten die nooit erg grote aantallen bereiken, of waarvan de biomassa's laag liggen minder "gediscrimineerd". Werken met cutlevel 0 is hiervan een extreem voorbeeld : de aantallen op zich spelen niet de minste rol bij de analyse. Enkel de aan- of afwezigheid van een soort bepaalt de toekenning tot een bepaald cluster.

* Een andere benadering zou natuurlijk zijn om per soort een frequentie distributie op te stellen en op basis daarvan cutlevels per soort te berekenen. Dit zou aan elke soort een gelijk gewicht geven en bovendien toelaten om de verdeling over de verschillende cutlevels te maximaliseren en zo de klassificatie te optimaliseren.

* De hierboven beschreven werkwijze werd wel geïmplementeerd voor alle soorten samen : door een frequentie-distributie op te stellen voor alle soorten en monsters samen, krijgt men een beeld van de "gerealiseerde" aantallen in de dataset en wordt het mogelijk om de cutlevels zo maximaal mogelijk te spreiden over het aantal positieve waarnemingen (positief wil zeggen dat de dichtheid of biomassa groter is dan 0). Op die manier worden de pseudospecies optimaal gedifferentiëerd en komen ze evenveel voor. Ook de two-way tabel wordt op die manier veel beter interpreteerbaar.

* Nog een andere mogelijkheid vormen de zelf toe te kennen gewichten, die binnen elke soort worden vermenigvuldigd met de gewichten van de pseudospecies (Ter Braak,1982). ARENICOLA marina zou bijvoorbeeld een "gewicht" kunnen toegekend worden gelijk aan de inverse van het gemiddeld lichaamsgewicht (binnen het studiegebied of berekend op basis van literatuur- gegevens), en zou daarmee op gelijke voet komen te staan met bijvoorbeeld BATHYPOREIA elegans, die nooit erg hoge biomassa's bereikt. Dezelfde redenering zou natuurlijk kunnen doorgetrokken worden naar dichtheden. In feite vormt de hier voorgestelde techniek een mijns inziens ekologisch beter gefundeerd alternatief voor de transformatie (relativering) van de data. Is men geïnteresseerd in een

klassificatie van monsters in de hoop een beter inzicht te krijgen in de verdeling van organismen van verschillende trofische niveaus in functie van het voedselaanbod, dan hebben de hierboven beschreven strategieën uiteraard weinig zin en werkt met men beter met de ruwe biomassa's.
Naast de soorten kan ook aan de monsters een gewicht gegeven worden.

De default cutlevels (0 2 5 10 20), zoals oorspronkelijk voorzien door Hill (1979) zijn zeker niet geschikt voor niet getransformeerde data. Ze werden ontworpen voor het werken met bedekkingspercentages (0-100%), waarvan de som per monster (opname) wel de 100% mag overschrijden, en zijn bijgevolg ook geschikt voor gestandardiseerde of gerelativeerde data, maar zeker niet voor ruwe data. De dichtheden over alle soorten en stations binnen één enkele dataset zijn invers exponentieel verdeeld: door de cutlevels ongeveer lineair te kiezen krijgen relatief weinig soorten, die meer dan gemiddeld abundant zijn, een zeer hoog gewicht en de vele weinig abundante soorten veel te weinig (discriminerende) invloed op de klassificatie. Ons lijken daarom log normale cutlevels (0 1 2 4 8 16..) ekologisch meest zinvol en hierin werd dan ook voorzien d.m.v. een extra default optie in TWINSPAN.

DISCRIM

op basis van hiërarchische klassificatie van de stations, zoals die verkregen werd met TWINSPAN (of evt. andere technieken), een klassificatie maken van de milieufactoren en proberen "indicator milieufactoren" te onderscheiden.

EEN UITGEWERKT VOORBEELD (nader uit te werken)

Teneinde de hier besproken methoden te illustreren werken we een simpel theoretisch voorbeeldje volledig uit. Waar nodig, om bepaalde aspecten te illustreren, worden voorbeelden van diverse andere (reële) datasets gebruikt.

De data

De hypothetische data zijn afkomstig van een bemonstering van de bodemfauna van de slikken gelegen in een bepaald estuarium. Op 12 punten werd de fauna bemonsterd en werden gegevens ivm enkele abiotische factoren verzameld. De datamatrix is samengevat in tabel . In totaal werden 14 soorten aangetroffen. Gegevens over het sediment (hoeveelheid kalk in de bodem, Ph van het interstitieel water en slib of zandbodem) en het bovenstaande water (zuurstofverzadiging, saliniteit en gehalte aan zware metalen) zijn opgenomen in tabel x.

voorbeeld van een frequentiedistributie van enkele soorten
listing van dataedit
transformaties
PCA
RA
Dca
verschillende similariteits indices

groeperingen
Twinspan
Canoco
diversi

MDSCAL -

CONCLUSIE

DANKWOORD

Dit rapport kon tot stand komen dank zij de financiele hulp van Rijkswaterstaat, Dienst Getijdenwateren (Nederland) en het Instituut voor Natuurbehoud (België). Een eerste faze van het rapport werd voorbereid door Dirk Develter. iets over de programma's
lezers

LITTERATUUR

Alvey, N. G. et al. 1980. GENSTAT. A general statistical program. Numerical algorithms group, Oxford.

Austin, M. P., 1976. On non linear species models in ordination. Vegetatio, 33: 33-41.

Clifford, H. T. & Stephenson, W., 1975. An Introduction to numerical classification. Academic Press, New York.

Conover, W. J., 1980. Practical Nonparametric Statistics. Tweede uitgave, J. Wiley, New York.

Cochran, W. G., 1967. Sampling Techniques, Wiley, New York.

Dixon, W. J. (Ed.), 1983. BMDP Statistical Software. Univ. of california Press, Berkely.

Gauch, H. G. 1982. Multivariate analysis in community Ecology. Cambridge University Press, Cambridge.

Gittins, R., 1969. The application of ordination techniques. In: Rorison, I. H. (Ed.), Ecological aspects of the mineral nutrition of plants, p37-66 Blackwell, Oxford.

Greig-Smith, P., 1980. The development of numerical classification and ordination. Vegetatio, 42: 1-9.

Greig-Smith, P., 1983. Quantitative plant ecology. Studies in Ecology, Vol. 9. Derde uitgave, Blackwell Scientific Publications, Oxford.

Heip, C., Herman, P. M. J. & Soetaert, K., 1988. Data processing, evaluation and analysis. IN ????

Heiser, W. J., 1981. Unfolding analysis of proximity data. Thesis. University of Leiden, 273 pp.

Hermy, M., 1984. The creation and the analysis of data matrices in vegetation science. Bull. Soc. Roy. Bot. Belg. 117: 63-72.

Hermy, M., 1985. Ecologie en fytosociologie van oude en jonge bossen in Binnen-Vlaanderen. Doktoraatsverhandeling RUG.

Hill, M. O. 1979a. DECORANA - A FORTRAN program for detrended correspondance analysis and reciprocal averaging. Cornell Univeristy, Ithaca, N. Y., 52pp.

Hill, M. O. 1979b. TWINSPAN - A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of individuals and attributes. Cornell Univeristy, Ithaca, N. Y., 90pp.

Hill, M. O., Bunce, R. G. H. & Shaw, M. W., 1975. Indicator species analysis, a divisive polythetic method of classification and its application to a survey of native pinewoods in Scotland. Journal of Ecology 63: 597-613.

Jongman, R. H. G., Ter Braak, C. J. F., & Van Tongeren, O. F. R., 1987. Data analysis in community and landscape ecology. PUDOC, Wageningen.

Legendre, L. & Legendre, P., 1986. Ecologie Numerique. Tome 1 en Tome 2. Tweede uitgave, Masson, Les pressses de l'Université du Quebec, Quebec.

Manly, B. F. J., 1986. Multivariate statistical methods. A primer. Chapman and Hall, New York.

Mohler, C. L., 1987. COMPOSE. A program for formatting and editing data matrices. Microcomputer power, Ithaca, New York.

Pianka, E. R., 1987. The subtlety, complexity and importance of population interactions when more than two species are invloved. Revista Chilena de historia Natural 60: 351-361.

Pielou, E. C., 1984. The interpretation of ecological data. A primer on classification and ordination. Wiley & Sons, New York.

SAS, 1982. SAS Users's guide: statistics. SAS Institute Inc., Cary, N.C.

Siegel, S., 1956. Nonparametric Statistics for the behavioral sciences. McGraw-Hill Kogakusha, Tokyo.

Sneath, P. H. A. & Sokal, R. R., 1973. Numerical taxonomy. Freeman, San Fransico.

Snedecor, G. W. & Cochran, W. G., 1980. Statistical methods. Zevende uitgave, Iowa State University Press, Ames, Iowa.

Sokal, R. R., & Rohlf, F. J., 1981. Biometry, The principles and practice of statistics in biological research. Tweede uitgave, Freeman and Company, New York.

SPSS, 1986. SPSS-X User's guide. Tweede uitgave, SPSS Inc., Chicago

Ter Braak, C. J. F., 1986a. CANOCO - a Fortran programm for CANOnical Community Ordination by (partial) (detrended) (canonical) correspondance analysis, principal components analysis and redundency analysis (version 2.0). Manual, TNO, Wageningen.

Ter Braak, C. J. F., 1986b. Canonical correspondance analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67: 1167-1179.

Ter Braak, C. J. F., 1987a. The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio 69: 69-77.

Ter Braak, C. J. F., 1987b. Ordination. In Jongman, R. H. G., Ter Braak, C. J. F., & Van Tongeren, O. F. R. (Eds.). Data analysis in community and landscape ecology. pp. 91-173, PUDOC, Wageningen.

Wishart, D., 1978. CLUSTAN user manual. Progr. Library Unit. Edinburgh Univ. Press, Edinburgh.